COMMUTATOR-RESIDUAL QUANTUM CIRCUITS: MITIGATING BARREN PLATEAUS WITH AN EXPONENTIALLY LARGE LIE ALGEBRA

Anonymous authors
Paper under double-blind review

ABSTRACT

Barren plateaus, the exponential vanishing of gradient variances in parametrized quantum circuits, are the central obstacle to scaling variational algorithms. We introduce the *commutator-residual* ansatz, a circuit of two-qubit group commutators whose generators are symmetrically drawn from a Y-odd Pauli pool. The pool's dynamical Lie algebra (DLA) is the Y-odd sector of $\mathfrak{su}(2^n)$, dimension $(4^n - 2^n)/2$, and the closed form exposes a commutator-degeneracy trap that independent sampling falls into and the CoPairSite protocol, which pairs anti-commuting generators on one site, eliminates. A cross-driving lock ties every gradient to a single-qubit Y expectation, and this carrier mechanism motivates a conjectured polynomial lower bound in the small-initialization regime, which would explain why an exponentially large DLA need not imply a barren plateau. At $n = 16$ the CoPairSite variance exceeds the hardware-efficient ansatz (HEA) by three orders of magnitude and decays with width at a fitted log-log slope near -1 against HEA's -8; the surviving gradients carry through to faster convergence and lower final errors on ground-state preparation and phase-recognition tasks, and the variance contrast survives depolarizing noise, a global cost, and identity-block initialization.

1 INTRODUCTION

Variational quantum algorithms (VQAs) are the dominant paradigm for extracting value from near-term quantum processors (Peruzzo et al., 2014; McClean et al., 2016; Cerezo et al., 2021a; Bharti et al., 2022), where a parameterized quantum circuit (PQC) prepares a state, a classical optimizer tunes the parameters, and parameter-shift or adjoint rules supply gradients (Schuld et al., 2019). The premise stands or falls on whether those gradients survive as the circuit grows.

The answer is often no. Barren plateaus (BPs), identified by McClean et al. (2018), are initialization ensembles in which gradient variances vanish exponentially in the qubit count n or the depth L, so an exponentially large number of shots is needed per gradient step. The causes are well mapped, from expressibility and 2-design proximity to cost locality, entanglement spread, and noise (McClean et al., 2018; Holmes et al., 2022; Cerezo et al., 2021b; Ortiz Marrero et al., 2021; Patti et al., 2021; Wang et al., 2021); the sharpest handle is the dynamical Lie algebra (DLA), the operator space generated by repeatedly commuting the circuit's elementary gates (Larocca et al., 2022; Ragone et al., 2024; Fontana et al., 2024; Anschuetz, 2025). Gradient variance is governed by the DLA dimension together with the input state, the observable, and the sampling measure on the Lie group (Ragone et al., 2024; Fontana et al., 2024). For ensembles that approximately cover that group, variances scale with $1/\dim \mathfrak{g}$, so polynomial-DLA circuits escape BPs while an exponentially large DLA signals danger unless the ensemble concentrates away from group coverage. The concentration breaks gradient-free optimizers (Arrasmith et al., 2021; 2022) and precludes learning scramblers (Holmes et al., 2022).

Most mitigation strategies leave the generator class untouched. Initialization reshapes the parameter distribution (Grant et al., 2021; Liu et al., 2021; Zhang et al., 2022; Mele et al., 2022; Park & Killoran, 2024; Park et al., 2024), optimization intervenes during training (Skolik et al., 2021; Volkoff & Coles, 2021; Sack et al., 2022; Grimsley et al., 2023), and symmetry-restricted ansätze

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shrink the DLA by construction (Yan et al., 2024; Sauvage et al., 2024; Park & Killoran, 2024). Measurement-based circuits curb the entanglement growth behind the plateau by inserting projective measurements (Wiersema et al., 2023). All share one implicit assumption, that the generator class is fixed and generic, so the DLA is either the full algebra or a problem-symmetry sector and trainability must be recovered around it. Three facts put that assumption under pressure. Grant-style identity-block initialization repairs the hardware-efficient ansatz (HEA) at shallow depth but the repair decays with depth and width. Residual and echo constructions that do start from the generator class (Kashif & Al-Kuwari, 2024; Khatun et al., 2026) suffer an inversion cost linear in depth. And the $O(1/\dim)$ bounds presuppose ensembles that approximately cover the Lie group, yet a commutator-built family with an *exponentially* large DLA shows measured variances that stay near depth-flat over the tested widths and exceed the $1/\dim$ prediction by five to six orders of magnitude, because the ensemble concentrates near the identity. The coverage premise therefore directed effort toward shrinking the algebra, and a generic pool's DLA can only be bounded, not computed, so the per-unit statistics of commutator sampling stay out of reach. *Can the generator class itself be built so that the ensemble concentrates near the identity, keeping the gradients flat while the algebra stays exponential?*

We propose the *commutator-residual* ansatz, whose exact Lie-algebraic structure removes both barriers. The closed-form DLA makes the trap and the escape measurable, and the group-commutator architecture turns the concentration near the identity from a weakness into the mechanism behind depth-flat variance. Each layer is a chain of two-qubit *group commutators*, each unit applying its generator pair forward and then backward, the two generators drawn *symmetrically* from a single Pauli pool. The unit is exactly the identity whenever $\theta = 0$ *or* $\phi = 0$, so the circuit starts at the identity and every layer is a second-order evolution with effective generator $[A, B]$; a unit's inverse is a constant-depth four-evolution circuit, with no inversion of an $O(\text{depth})$ accumulated block as in residual constructions. Three results support this design, one algebraic, one mechanistic, and one at the task level.

- We prove that the Y-odd pool has a DLA exactly equal to the Y-odd sector of $\mathfrak{su}(2^n)$, dimension $(4^n - 2^n)/2$, half the full algebra and yet exponentially large. The closed form exposes a commutator-degeneracy trap that no bounded-DLA analysis can see, since independent two-generator sampling wastes the parameter budget on identity units, and the CoPairSite protocol eliminates it.
- We conjecture the lower bound $\mathrm{Var}(\partial_\theta C) \geq c\,\sigma^6/\mathrm{poly}(n, L)$ for the parameters nearest the cost site in the small-σ regime, and we prove the exact mechanism behind it. The mechanism rests on the cross-driving lock, which expands each gradient as a $\sin(2\phi)$-modulated single-qubit Y expectation with linearly growing operator weight, giving a first-principles account of why an exponential DLA need not force a barren plateau; at $\sigma = O(1)$ the same lock explains the measured depth-flat variance.
- We demonstrate on the transverse-field Ising model (TFIM) with the variational quantum eigensolver (VQE) that CoPairSite gradients stay an order of magnitude larger than HEA's throughout training and converge where HEA stalls, that the advantage survives controlled parameter and gate budgets, and that the residual mechanism stays active while the cost operator remains confined to an exponentially large parity sector rather than a simulable subspace.

2 The Commutator-Residual Ansatz

2.1 Commutator units and pools

Let $A_{l,i}, B_{l,i}$ be Pauli strings acting on adjacent pairs of an n-qubit chain. Each layer consists of $n - 1$ units with independently drawn chain edges; for each unit,

$$U_l \;=\; \prod_{i=1}^{n-1} \underbrace{e^{i\theta_{l,i}A_{l,i}}\, e^{i\phi_{l,i}B_{l,i}}\, e^{-i\theta_{l,i}A_{l,i}}\, e^{-i\phi_{l,i}B_{l,i}}}_{\text{group commutator unit } R_{l,i}}, \qquad U_{L:1} = U_L \cdots U_1. \tag{1}$$

Each unit has exactly two parameters. By the Baker-Campbell-Hausdorff (BCH) expansion (Wilcox, 1967)

$$\log R_{l,i} \;=\; -\theta_{l,i}\phi_{l,i}\,[A_{l,i}, B_{l,i}] \;+\; O\big(\theta^2\phi,\, \theta\phi^2\big), \tag{2}$$

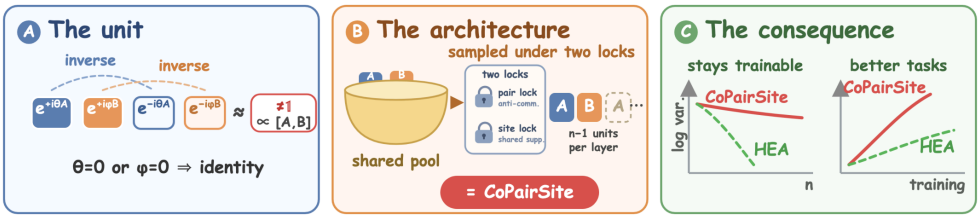


Figure 1: Method overview. *Panel A*: the group commutator unit, an ordered product of two inverse gate pairs that collapses to the identity when either parameter vanishes and otherwise evolves along $[A, B]$. *Panel B*: each layer chains $n-1$ units; the generator pair (A, B) is freshly sampled per unit from one shared pool, under an anti-commuting draw (pair lock) and a shared support (site lock) that together define CoPairSite. *Panel C*: the consequence in sketch form, log variance stays near-flat in n while HEA decays, and task progress follows during training.

the unit is a *second-order* evolution whose effective generator is the commutator $[A, B]$. Vanishing of either parameter collapses the unit to the identity, and at small parameter scale the unit is a rotation along $[A, B]$ with strength governed by the product $\theta\phi$. Figure 1 sketches the unit and layer layout.

The ansatz is parameterized by Pauli pools. The *unconstrained* pool $\mathcal{P}_{\mathrm{u}}$ contains all nine two-qubit Pauli strings without identity factors; the Y*-odd* pool $\mathcal{P}_{\mathrm{y}} = \{XY, YX, YZ, ZY\}$ (odd Y count) samples four of them. The commutator of two Y-odd strings remains Y-odd, so the sector they generate closes with a closed-form dimension; which letter carries the grading is a convention, the three odd pools being equivalent under single-qubit Clifford relabelings. The *sector* $\mathfrak{g}_{\mathrm{y}}$, the real span of *all* Y-odd Pauli strings on n qubits (including YI, IY), strictly contains the pool, since $XY \cdot ZY \propto YI$ enters the dynamics although YI is never sampled.

2.2 The commutator-degeneracy trap and CoPairSite sampling

Degeneracy enters at two levels. (i) At the string level, four of the six unordered Y-odd string pairs commute, so the unit is an identity whenever $[A, B] = 0$; (ii) more severely, A and B on disjoint qubits commute *regardless of strings*, and disjoint supports dominate at large n under Naive's independent position sampling. Monte Carlo over 2×10^5 samples per configuration shows Naive spends most of its parameter budget on identity units, for both pools, a general trap for commutator-type ansätze with independent two-generator sampling. Each identity unit carries parameters but contributes no generator, so the waste shows up only in the gradients.

The two locks target the two degeneracy levels. The *pair lock* restricts (A, B) to the anti-commuting pairs $\{XY, ZY\}$, $\{YX, YZ\}$, removing string-level degeneracy; the *site lock* forces A, B onto the same adjacent pair, removing the dominant support-level degeneracy. We evaluate the 2×2 family they define. *Naive* applies neither; *CoPair* only the pair lock; *CoSite* only the site lock; *CoPairSite* both, removing all degeneracy so every unit is active. Full-precision identity rates and the four-protocol illustration are in Appendix F.

3 Theoretical Analysis

With the ansatz fixed, this section establishes the two properties that independent sampling cannot deliver simultaneously, the algebraic reach of the pool and depth-robust gradients; proofs and machine-verification details are collected in the appendices, and the main text states the results and the mechanism.

3.1 Exact DLA structure

The exact DLA structure follows from one parity fact. For Pauli strings P, Q, let $y(P)$ denote the number of Y factors in P; multiplying strings shifts the parity as $y(PQ) \equiv y(P) + y(Q) + \mathbf{1}[P, Q \text{ anticommute}] \pmod 2$, so nonvanishing commutators of Y-odd strings are again Y-odd.

Theorem 1 (DLA of the Y-odd commutator pool)**.** *Let $\mathfrak{g}_{\mathrm{y}}$ be the dynamical Lie algebra generated by the generators of pool $\mathcal{P}_{\mathrm{y}}$ placed on all adjacent pairs of an n-qubit chain, $n \geq 2$. Then*

$$\mathfrak{g}_{\mathrm{y}} \;=\; \mathfrak{g}_{\mathrm{y}} \;=\; \mathrm{span}_{\mathbb{R}}\big\{\, iP \;:\; P \textit{ Pauli string},\; y(P) \textit{ odd} \,\big\}, \qquad \dim \mathfrak{g}_{\mathrm{y}} = \frac{4^n - 2^n}{2}. \tag{3}$$

The unconstrained pool and the standard R_y–R_z-with-entangling-layer hardware-efficient ansatz (HEA) generate the full $\mathfrak{su}(2^n)$, $\dim = 4^n - 1$.

Each unit of Eq. (1) contains $e^{i\theta A}$ and $e^{i\phi B}$ as literal evolutions, so the pool strings belong to the gate set of the ansatz itself and Theorem 1 characterizes that ansatz, not a surrounding library; the leading-order generator $\theta\phi[A, B]$ of a near-identity unit is a separate object, taken up below. The Y-odd constraint keeps half the algebra, an exponentially large vocabulary reduced only by a constant fraction. In the Lie-algebraic theory of barren plateaus (Ragone et al., 2024; Fontana et al., 2024) the same dimension enters as a danger. Gradients uniform over the algebra have variance $\Theta(1/\dim \mathfrak{g})$, so an exponential DLA seems to condemn any ansatz to a plateau. CoPairSite keeps the exponential vocabulary while its measured gradients hold up where the $1/\dim \mathfrak{g}$ reading predicts decay, so algebraic reach and initialization-time gradient concentration come apart rather than trade; what the unit pays for is the size of the generator family, not the reach of the algebra.

3.2 Gradient duality and depth robustness

We make the separation exact by writing the gradients out explicitly rather than reading them off a shift rule, which does not apply to commutator units. Write the unit as $U = WV$ with $W = G_1G_2 = e^{i\theta A}e^{i\phi B}$, $V = G_3G_4 = e^{-i\theta A}e^{-i\phi B}$, and $O' = W^\dagger O W$. Two identities are used below. An *exact gradient duality*, under which each driving term contains only the *other* parameter, and a *four-point parameter-shift rule*, which exploits the frequency support $\{0, \pm 2, \pm 4\}$ of $C(\theta)$ so that four evaluations at fixed shifts $\{\pi/4, \pi/8\}$ recover the gradient exactly. Both hold at machine precision; the generator-set completeness and the remaining proofs are in Appendices H and G.

Lemma 1 (Cross-driving locks each gradient onto a single-qubit Y)**.** *For every anti-commuting pair of the CoPairSite pool the θ-driving term expands exactly as*

$$e^{-i\phi B} A\, e^{i\phi B} - A = -2\sin^2\phi\; A \pm \sin(2\phi)\, Y, \tag{4}$$

where Y is the single-qubit Pauli-Y on the site at which A and B disagree (the X/Z site; the pair coincides on the shared-Y site), and the sign is the pair orientation. Inserting (4) *into the duality identity for $\partial_\theta C$ gives the exact decomposition*

$$\frac{\partial C}{\partial \theta} = \pm 2\sin(2\phi)\,\mathrm{Im}\,\langle 0|V^\dagger Y O' V|0\rangle - 4\sin^2\phi\,\mathrm{Im}\,\langle 0|V^\dagger A O' V|0\rangle. \tag{5}$$

The ϕ-gradient admits the same single-qubit structure with a single factor,

$$\frac{\partial C}{\partial \phi} = \pm 4\sin\theta\;\mathrm{Im}\,\langle 0|G_4^\dagger Y O' V|0\rangle, \qquad G_4 = e^{-i\phi B}, \tag{6}$$

with the same X/Z-site single-qubit Y. Both gradients are carried by a single-qubit Y expectation modulated by the other *parameter.*

The value of (5) is what it reveals, reducing the variance question to the fate of one local observable under the circuit's conjugations.

3.3 The variance lower bound

The lock of Lemma 1 fixes the gradient of every unit, but a variance bound must cover a whole chain, so the open question is what L layers of conjugation do to the single-qubit carrier. Each unit adds at most one site of support, the carrier's weight grows linearly in depth, and the carrier then overlaps the local cost along polynomially many positions. This conjugation control is the mechanism behind a conjectured polynomial lower bound in the small-σ regime. We state the bound as a conjecture rather than a theorem, because three ensemble steps resist a complete proof; the all-parameters strengthening does not hold, for a structural reason given below.

Conjecture 1 (Small-σ variance scaling)**.** *Let $C = \langle 0^n|U^\dagger Z_0 U|0^n\rangle$, where U is an L-layer chain* (1) *of CoPairSite units on n qubits, each unit built from an anti-commuting pair (A, B) of Y-odd strings on a shared adjacent position, and let the parameters be drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$ with $\sigma \le \sigma_0$ for a sufficiently small constant σ_0. Write $d \in \{0, \dots, n-1\}$ for the distance from a parameter's carrier site to the cost site 0 along the chain. (i) For the parameters of the unit adjacent to the cost site, $d = 0$,*

$$\mathrm{Var}[\partial_\theta C] \;\ge\; \frac{c\,\sigma^6}{\mathrm{poly}(n, L)}\,, \tag{7}$$

for a constant $c > 0$, and the same bound holds for $\mathrm{Var}[\partial_\phi C]$, so no exponential decay occurs for these parameters in this regime. (ii) The variance scale degrades with the carrier–cost distance d, and a uniform σ^6 bound over all parameters is not expected to hold.

The exact carrier and support lemmas prove the carrier and its one-site support growth; the conjectured step is the ensemble statement, a polynomial fraction of configurations keeping the carrier matched to the cost with a non-cancelling coefficient. The reflection symmetry that would force the mean to vanish does not exist for the ordered product, and the estimates of Napp (2022) bound reachability from above, not the overlap fraction from below. The distance dependence in part (ii) comes from a support-matching count. A carrier at chain distance d contributes an expected gradient of order σ^{3+d} and a variance of order σ^{6+2d}, so the uniform σ^6 form can hold only at $d = 0$. The full argument is in Appendix A.

The design claim is not that the DLA is small, but that its size is not paid for in gradient variance. (5) shows why the commutator gradient escapes the $1/\dim \mathfrak{g}$ scaling. Its leading term is a single-qubit Y expectation regenerated at every unit rather than a uniform sample over the algebra. A channel-level diagnostic agrees. The averaged channel $\Lambda_L = \mathbb{E}[U_{L:1}^\dagger(\cdot)U_{L:1}]$ contracts Y-odd operators almost completely by $L = 40$, yet cost-relevant retention decays only sub-exponentially, $0.96 \to 0.31$ over $L = 1 \to 40$ at $\sigma = 0.3$, a strongly *direction-selective* contraction consistent with regeneration.

Numerical status of the bound. The conjectured σ^6 order is checked at $d = 0$ by exact state-vector computation at $n = 2$, the largest system admitting a decade-wide σ scan; the admissible σ_0 is at least 0.18, and at fixed small σ the variance *rises* with depth rather than decaying. The tasks run at $\sigma = O(1)$, outside the conjecture's scope, but the carrier lock is exact for every σ, so the carrier and its linear-in-L weight growth hold verbatim there. The σ-grid, seed counts, and fitted slopes are in Appendix B.

Conjecture 2 (Task-regime lower bound)**.** *For the CoPairSite chain at initialization scale $\sigma = O(1)$ and local or gapped cost observables, there exists $c(\sigma) > 0$, bounded away from zero on compact σ-intervals, such that $\mathrm{Var}[\partial_\theta C] \ge c(\sigma)/\mathrm{poly}(n, L)$ for all L.*

Proven are the carrier mechanism (Lemma 1) and its one-site support growth; conjectured are the two bounds above, which the measured exponents support up to $n = 16$, with no cell near the exponential scale.

4 Experiments

The claim under test is architectural, so the experiments hold everything but the architecture fixed. All variants (HEA, Naive Y-odd commutator, CoPairSite) run in one implementation with identical simulators, optimizers, seeds, and data; HEA uses $R_y R_z$ layers with a CNOT ring, commutator variants use layout (1). The initialization scale σ, the width of the parameter distribution $\mathcal{N}(0, \sigma^2)$, is an architecture-independent treatment variable shared by every variant, swept over $\{0.1, 0.3, 0.5, 0.7\}$ from narrow to near-random starts. Throughout, the reported gradient quantity is the mean-squared-gradient signal, the per-parameter expectation of g_j^2 averaged over initializations; the strict per-parameter variance agrees with it to within 3% in every tested configuration. The section proceeds as a sequence of validity checks, each block ruling out one alternative explanation of the one before it.

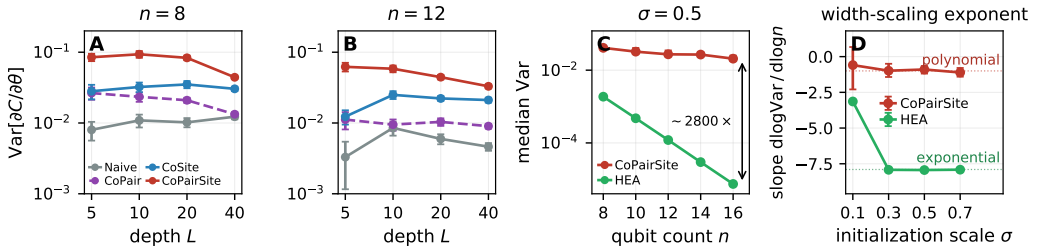


Figure 2: Protocol verdict and width scaling at $L = 40$ (50 initializations per point). *Panels A, B:* verdict experiment at n=8 and n=12, $\sigma = 0.5$ (standard-error bars, CoPair dashed); all four sampling protocols remain depth-flat, fully active CoPairSite sits 1–1.5 orders of magnitude above Naive. *Panel C:* $\sigma = 0.5$, absolute median variance against qubit count with bootstrap intervals; the exponents are fitted against the actual qubit counts. *Panel D:* the fitted exponent against initialization scale; a flat exponent means polynomial decay in the width, a steep one signals exponential decay.

4.1 Protocol verdict and n-scaling head-to-head

Naive's depth-flat variance could be an artifact of its 80–90% identity units (a sparse, effectively shallow circuit trivially avoids deep-circuit decay). A four-protocol experiment disentangles the two candidate mechanisms, sparsity and structural flatness; only the (A, B) sampling rule changes, everything else (gate structure, parameterization, initialization $\theta, \phi \sim \mathcal{N}(0, \sigma)$, cost Z_0, 50 random initializations) is held fixed.

Deep-circuit variance ensembles are heavy-tailed, and Naive and CoSite additionally spike *structural zeros* (circuits whose every unit is an identity), so the verdict comparison uses medians over nonzero seeds; the all-seed mean preserves the ordering. Three conclusions follow (Figure 2 panels A, B). Depth-flatness is structural. fully active CoPairSite retains flat variance with *no* structural zeros at any depth, while Naive's zero fraction decays from 74% at $L = 5$ to 0% at $L = 40$, ruling out sparsity. The correction also removes the absolute-signal penalty. At $n = 12$, $L = 40$, $\sigma = 0.5$, CoPairSite reaches 3.3×10^{-2} against Naive 4.6×10^{-3} and HEA 1.19×10^{-4}. Position locking alone recovers most of the signal, pair locking alone much less.

With the measurement validated, the same protocol scales to $n = 14, 16$, where CoPairSite's depth-flatness persists with no structural zeros. The decisive comparison is the *n-scaling* against HEA at matched depth and initialization (Figure 2 panels C, D, median over 50 initializations, $L = 40$). At $\sigma = 0.5$, HEA collapses from 1.9×10^{-3} at $n = 8$ to 7.4×10^{-6} at $n = 16$ while CoPairSite decays only from 4.1×10^{-2} to 2.1×10^{-2}, fitted exponents -7.9 and -0.9. Panel D reports the exponent at $\sigma \in \{0.1, 0.3, 0.5, 0.7\}$ with bootstrap intervals over the seeds. HEA reaches -7.9 for $\sigma \geq 0.3$ and CoPairSite stays between -0.6 and -1.1 throughout, a gap of about seven exponent units stable from $\sigma = 0.3$ to 0.7. At $\sigma = 0.1$ HEA has not yet entered its steep regime, so that point lies inside the small-σ regime of Conjecture 1 ($\sigma_0 \geq 0.18$). At $\sigma = 0.5$, $n = 16$, CoPairSite exceeds HEA by $\sim$2800×.

4.2 VQE ground-state preparation

The measurement question is closed, but flat variance alone does not certify utility, so we test two uses, minimizing a fixed Hamiltonian's energy and learning a decision rule from data.

We first consider the TFIM ($n = 8$, $h = 0.5$, $L = 12$, $E_0 = -8.509$, Adam, 150 steps, 10 seeds; Figure 3a). Final energies (mean±s.d.) show HEA splitting at $\sigma = 0.5$ (-7.36 ± 0.79) and collapsing at $\sigma = 0.7$ (-6.85 ± 0.08); Naive converges at every σ but falls short of E_0 by $\sim$0.05–0.08; CoPairSite converges to gap ≤ 0.01 at *every* combination of σ and seed.

The in-task gradients close the loop with the variance measurements. At $n = 12$ ($L = 12$, $\sigma = 0.5$, 150 Adam steps, 10 seeds; Figure 3b), HEA's gradients stay at $\|\nabla C\|_{\text{rms}} \sim 8.9 \times 10^{-3}$ and the energy stalls near -2.07; CoPairSite maintains gradients $\sim$20× larger (1.78×10^{-1}) and converges to $E_0 + 0.008$ ($E_0 = -12.763$); Naive sits between (4.0×10^{-2}, gap 0.22). The CoPairSite gradient stays above HEA's at every one of the 150 steps, so the surviving gradients persist through optimization rather than holding at initialization alone. The Heisenberg model ($n = 8$, antiferro-

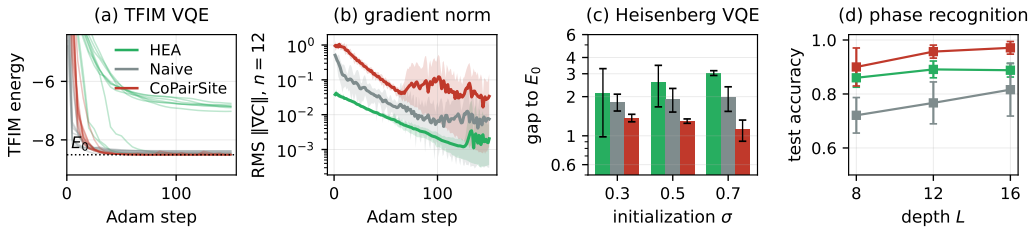


Figure 3: Task evidence in one implementation. (a) TFIM VQE trajectories at $\sigma = 0.5$, $n = 8$ (10 seeds). (b) In-task gradient norms at $n = 12$, $\sigma = 0.5$ (min–max bands over 10 seeds, thick lines means). (c) Heisenberg final-energy gaps (10 seeds). (d) TFIM phase-recognition test accuracy (10 seeds, 100 test states).

Table 1: Baseline comparison of the final energy gap to E_0 on TFIM VQE (n=8, L=12, 150 Adam steps, 10 seeds, per-seed median). Heavy-tailed seed distributions, so the median tracks the typical seed; the last column counts seeds converging to gap <0.1 out of 30. All baselines follow their sources with the same optimizer budget. P is the parameter count; the budget-matched comparison in Appendix D controls for the two-qubit gate budget.

Methods	P	$\sigma = 0.3$	$\sigma = 0.5$	$\sigma = 0.7$	Converged
HEA (random init)	192	0.007	1.616	1.701	12/30
HVA (symmetry-restricted)	192	0.010	0.019	0.009	27/30
Layerwise HEA	192	0.017	1.835	1.843	10/30
Identity-block HEA	192	0.004	0.005	0.012	28/30
CoPairSite (ours)	**168**	**0.002**	**0.003**	**0.008**	**30/30**

magnetic, $L = 12$, $E_0 = -14.604$; Figure 3c) is expression-limited for all methods, and CoPairSite is simultaneously the most stable and, at large initialization scale, the best, reaching -13.49 ± 0.22 at $\sigma = 0.7$ versus Naive -12.64 ± 0.44 and HEA -11.57 ± 0.14.

The contrast is architectural rather than an artifact of near-identity starts or cost locality. Identity-block initialization of HEA repairs shallow depth only and decays to random-init levels at $L = 40$, the global TFIM cost widens the margin to $18\times$, and matching on parameters and on two-qubit gates separately preserves the advantage, so the gap tracks neither the start nor the gate budget. What is conventional versus load-bearing, with both ablations, is in Appendix C.

Table 1 widens the comparison to existing mitigation strategies under the same budget, including a symmetry-restricted Hamiltonian variational ansatz (HVA), layerwise training, and identity-block initialization. Identity-block HEA is the strongest baseline, but at $\sigma = 0.7$ two of ten seeds fail catastrophically; layerwise training collapses at $\sigma \geq 0.5$; the HVA comes closest at $\sigma = 0.7$ but carries the Hamiltonian's generator structure by construction and has no analogue for the learning task below. What the comparison isolates is one fixed architecture that converges from every seed and holds the seed-failure rate at zero.

4.3 TFIM QUANTUM PHASE RECOGNITION

Binary classification of ferromagnetic vs. paramagnetic TFIM phases ($n = 8$; exact ground-state data perturbed by $\epsilon = 0.3\pi$ local rotations, 40 train / 100 test states, parity readout, $\sigma = 0.5$, 10 seeds) is nontrivial for a linear readout, since the unperturbed states have vanishing magnetization by symmetry. CoPairSite dominates at every depth (Figure 3d), reaching 0.971 ± 0.025 at $L = 16$, where Naive has plateaued (0.816 ± 0.103) and HEA trails (0.888 ± 0.027). The ordering matches the gradient evidence, classification progress tracks the size of the in-task gradients across the three variants.

4.4 NOISE ROBUSTNESS

What matters for the trained task is whether it survives noise, not only how the variance behaves. We run the TFIM VQE above with a local depolarizing channel on every qubit after every logical layer ($p = 0.005$); the CoPairSite/HEA variance margin over $p \in \{0, 0.002, 0.005, 0.01\}$ *widens* from

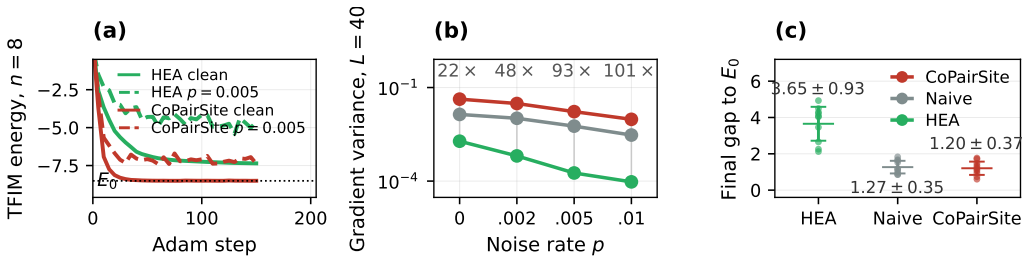


Figure 4: Task-level noise robustness under local depolarizing noise, the TFIM VQE of Section 4.2 (n=8, L=12, σ=0.5). (a) Mean energy trajectories at error probability $p = 0.005$ after every layer (10 seeds). (b) Gradient variance against the noise rate p at L=40 (20 seeds per point). (c) Final gap to E_0 per seed at $p = 0.005$ (10 seeds, mean±s.d.). The CoPairSite margin over HEA widens with p.

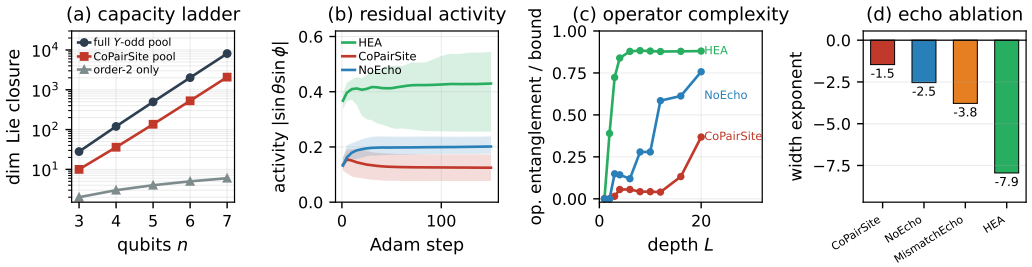


Figure 5: Capacity and mechanism controls. *(a)* Lie-closure dimension against width for the single anti-commuting family $\{XY, ZY\}$ that CoPairSite locks, the full Y-odd pool, and the order-2 Abelian direction, with closed forms drawn through the points. *(b)* Unit activity $|\sin\theta \sin\phi|$ during TFIM VQE training at $n = 8$, $\sigma = 0.5$ (10 seeds, bands min–max). *(c)* Operator entanglement of $U^\dagger Z_0 U$ against depth at $n = 6$, normalized by the half-cut bound. *(d)* Width-scaling exponent of the gradient variance for the echo ablations, $L = 40$, $\sigma = 0.5$.

22× to 101× (Figure 4b). Under the channel HEA stalls near gap 4 (3.65 ± 0.93, 10 seeds) while CoPairSite keeps descending to gap 1.20 ± 0.37 (Naive 1.27 ± 0.35), a 3× margin that mirrors the noiseless ordering. Gradients that survive the channel keep the energy decreasing where HEA's have vanished, a picture consistent with carrier regeneration at every unit (Figure 4a,c); full-precision variance tables are in Appendix I.

4.5 Capacity, Mechanism, and the Cost of the Echo

Four controls separate a usable design from an accounting trick (Figure 5). *Reach* (Figure 5a) is fixed by the construction, not by σ. One anti-commuting family, $\{XY, ZY\}$, closes on a Lie algebra of dimension $4^n/8 + 2^n/4$, 25.6% of the full Y-odd pool at $n = 7$ and still exponential. *Scale invariance* is measured on the exact Jacobian at $n = 5$, $L = 3$. The normalized spectrum moves by a cosine distance of only 0.012 over four decades of σ, whereas HEA's ratio falls by four orders of magnitude, so shrinking the residual rescales the reachable directions rather than deleting them. *Mechanism* survives training (Figure 5b). The per-unit activity $|\sin\theta \sin\phi|$ stays near 0.13 for CoPairSite over all 150 TFIM VQE steps while HEA's rises from 0.37 to 0.50, so the large gradient is not units saturating into unstructured rotations. *Scope* (Figure 5c) answers the simulability question sharply. Decomposing $U^\dagger Z_0 U$ at $n = 6$, the CoPairSite operator stays exactly inside the Y-even sector, 2078 of 2079 strings at $L = 12$ with no Y-odd component, while HEA fills the full 4095, and its operator entanglement still grows to 0.37 of the cut bound by $L = 20$ against HEA's 0.88. No-plateau here therefore implies neither simulability nor hardness.

The echo structure is load-bearing (Figure 5d). Replacing the matched inverse with a plain four-rotation block steepens the width exponent from -1.46 to -2.54, and an inverse drawn from a different anti-commuting pair is worse still, at -3.78, below the no-inverse variant and close to HEA's -7.95. What carries the depth-flat variance is the matching of the inverse to its forward generator, the condition that also collapses the unit to the identity when a single angle vanishes, not

the presence of an inverse. The no-inverse block starts above CoPairSite at shallow depth, 3.1×10^{-1} against 7.3×10^{-2} at $L = 5$, but that head start comes from its larger two-qubit pool and not from the echo structure. It decays by a factor of 6.8 from $L = 5$ to $L = 40$ and settles at CoPairSite's level there, its σ-scaling is nearly flat ($+0.7$ at $L = 3$ and 0 at $L = 10$) where CoPairSite holds near σ^6, and the activity and operator-entanglement diagnostics do not carry the same meaning for it. Exact closure dimensions, tangent spectra, and the full operator-complexity scan are in Appendix E.

5 Related Work

A pivotal line connects trainability to the DLA, reading the variance off the algebra's dimension and the ensemble's coverage of the Lie group (Larocca et al., 2022; Ragone et al., 2024; Fontana et al., 2024), via the adjoint representation (Fontana et al., 2024) and 2-local growth classifications (Wiersema et al., 2024); see Larocca et al. (2025) for a survey. Theorem 1 adds an exactly solvable member whose DLA is an explicit half-sector of $\mathfrak{su}(2^n)$. Our variances exceed the $1/\dim \mathfrak{g}$ prediction by 5–6 orders of magnitude (Section 4.1), a scope statement rather than a contradiction, since those bounds assume Lie-group-covering ensembles while the commutator unit concentrates near the identity (Eq. (2)) and never samples the whole algebra (Lemma 1).

Most mitigation strategies leave the generator class untouched, reshaping the parameter distribution or intervening during optimization (Grant et al., 2021; Zhang et al., 2022; Mele et al., 2022; Park & Killoran, 2024; Park et al., 2024; Skolik et al., 2021; Volkoff & Coles, 2021; Pesah et al., 2021; Sack et al., 2022; Patel et al., 2024); ours changes the generator algebra itself, holding a near-flat n-scaling where the identity-block reset no longer helps. Measurement-based circuits instead curb the plateau where it is driven, by inserting projective measurements that suppress entanglement growth, demonstrated for the Hamiltonian and hardware-efficient ansätze (Wiersema et al., 2023). The relief there is tied to conditioning on measurement outcomes, so shots split across branches and the mid-circuit back-action enters the noise model. Ours reaches the flat variance with a purely unitary circuit, no mid-circuit measurements, and a deterministic shot budget, and its reachable operator space stays in an exponentially large parity sector rather than collapsing toward a classically simulable one (Section 4.5). Symmetry-restricted ansätze shrink the DLA by construction (Yan et al., 2024; Sauvage et al., 2024) and tensor-network circuits (Cong et al., 2019) restrict it to a network-generated manifold; the bounded algebra that provable plateau-freedom asks for is also what invites classical simulation (Cerezo et al., 2023), whereas ours keeps the exponential vocabulary and owes its trainability to the ensemble. Quantum residual designs (Kashif & Al-Kuwari, 2024; Khatun et al., 2026) attach a trainable gate to an explicitly inverted block, whereas ours builds the inverse into the unit, so the inversion cost does not scale with depth. The trap of Section 2.2 is not one we have found described for commutator-type ansätze with independent two-generator sampling. The noise-induced plateau line (Wang et al., 2021) sets a second bar; our depolarizing measurements (Section 4.4) are the evidence that the concentration survives it.

6 Conclusion

We presented the commutator-residual ansatz, a design that pairs an exponential algebraic vocabulary with large initialization gradients at the widths we tested. The construction rests on three results, the closed-form Y-odd DLA (Theorem 1), the commutator-degeneracy trap (Section 2.2) that explains why naive sampling destroys gradients, and the pair-plus-site locks that defuse it by construction. Across TFIM and Heisenberg VQE and phase recognition, the surviving gradients yield faster convergence and lower final errors than HEA and naive sampling at matching budgets, the cost operator held throughout in an exponentially large parity sector. Replacing the matched inverse with a plain four-rotation block steepens the fitted width exponent from -1.5 to -2.5, so the echo inside each unit, not the circuit size, carries the flat variance, and the advantage widens from $22\times$ to $101\times$ under local depolarizing noise. The variance bound $\mathrm{Var}(\partial_\theta C) \geq c\,\sigma^6/\mathrm{poly}(n, L)$ and its task-regime extension remain conjectural, our evidence being a fitted width-scaling exponent near -1 up to $n = 16$ rather than a proof; proving the bound beyond its fitted regime and testing the locks on hardware-native generator sets are the natural next steps. An exponentially large dynamical Lie algebra is not a verdict on trainability, only a statement about which ensemble samples it.

Reproducibility statement

All experiments run in one pure-numpy state-vector simulator with adjoint-state gradients validated to $< 10^{-9}$; the algebraic and combinatorial statements are machine-verified at the stated precision, and the conjectured variance scaling is checked numerically in its regime (Appendices G–I). Each experiment has a reproducing script in the supplementary code.

AI use statement

We used generative AI tools for drafting text, implementing and debugging the simulators, and preparing figures. AI-suggested code was verified against independent analytical computations, and all AI-assisted text was reviewed by the authors, who take responsibility for the final content.

Ethics statement

This work is analytical and numerical. The experiments are state-vector simulations of synthetic spin models, with no human subjects, personal data, or external datasets beyond the released code archive. All authors have read the ICLR Code of Ethics and adhere to it, and we do not anticipate ethical concerns specific to this contribution.

References

Eric Ricardo Anschuetz. A unified theory of quantum neural network loss landscapes. In *International Conference on Learning Representations*, 2025.

Andrew Arrasmith, Marco Cerezo, Piotr Czarnik, Lukasz Cincio, and Patrick J. Coles. Effect of barren plateaus on gradient-free optimization. *Quantum*, 5:558, 2021.

Andrew Arrasmith, Zoë Holmes, Marco Cerezo, and Patrick J. Coles. Equivalence of quantum barren plateaus to cost concentration and narrow gorges. *Quantum Science and Technology*, 7: 045015, 2022.

Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94:015004, 2022.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3:625–644, 2021a.

Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12:1791, 2021b.

Marco Cerezo, Martín Larocca, Diego García-Martín, Nathan L. Diaz, Paolo Braccia, Enrico Fontana, Martin S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? or, why we need to rethink variational quantum computing. *arXiv preprint arXiv:2312.09121*, 2023.

Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15:1273–1278, 2019.

Enrico Fontana, Dylan Herman, Supratik Chakrabarti, Niraj Kumar, Romina Yalovetzky, Jamie Heredge, Sai Hari Sureshbabu, and Marco Pascuzzi. Characterizing barren plateaus in quantum ansätze with the adjoint representation. *Nature Communications*, 15:7171, 2024. doi: 10.1038/s41467-024-49910-w.

Edward Grant, Marcello Benedetti, Shuxiang Cao, Andrew Hallam, Joshua Lockhart, Vid Stojevic, Andrew G. Green, and Simone Severini. Initialization strategies for enhancing the performance of variational quantum optimizers. *Quantum*, 5:366, 2021.

Harper R. Grimsley, George S. Barron, Edwin Barnes, Sophia E. Economou, and Nicholas J. Mayhall. Adaptive, problem-tailored variational quantum eigensolver mitigates rough parameter landscapes and barren plateaus. *npj Quantum Information*, 9:19, 2023.

Zoë Holmes, Kunal Sharma, Marco Cerezo, and Patrick J. Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3:010313, 2022.

Muhammad Kashif and Shamma Al-Kuwari. Resqnets: a residual approach for mitigating barren plateaus in quantum neural networks. *EPJ Quantum Technology*, 11:32, 2024.

Anika Khatun et al. Hardware-efficient quantum residual neural networks without post-selection. *arXiv preprint arXiv:2604.06866*, 2026.

Martin Larocca, Nathan Ju, Danylo Guo, Patrick J. Coles, and Marco Cerezo. Diagnosing barren plateaus with tools from quantum optimal control. *Quantum*, 6:824, 2022.

Martín Larocca, Marco Cerezo, et al. Barren plateaus in variational quantum computing. *Nature Reviews Physics*, 2025. doi:10.1038/s42254-025-00813-9.

Hui-Ting Liu, Xuan Zhang, Hai-Long Zhu, Huan Wang, Cheng Shen, and Xiao Yuan. Escaping the barren plateau via covariance matrix adaptation. *arXiv preprint arXiv:2110.03631*, 2021.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18:023023, 2016.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9:4812, 2018.

Antonio A. Mele, Glen B. Mbeng, Giuseppe E. Santoro, Mario Collura, and Pietro Torta. Avoiding barren plateaus via transferability of smooth solutions in a hamiltonian variational ansatz. *Physical Review A*, 106:L060401, 2022.

John C. Napp. Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze. *arXiv preprint arXiv:2203.06174*, 2022.

Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement-induced barren plateaus. *PRX Quantum*, 2:040316, 2021.

Chae-Yeun Park and Nathan Killoran. Hamiltonian variational ansatz without barren plateaus. *Quantum*, 8:1239, 2024.

Chae-Yeun Park, Minhyeok Kang, and Joonsuk Huh. Hardware-efficient ansatz without barren plateaus in any depth. *arXiv preprint arXiv:2403.04844*, 2024.

Yash J. Patel, Akash Kundu, Mateusz Ostaszewski, Xavier Bonet-Monroig, Vedran Dunjko, and Onur Danaci. Curriculum reinforcement learning for quantum architecture search under hardware errors. In *International Conference on Learning Representations*, 2024.

Taylor L. Patti, Khadijeh Najafi, Xun Gao, and Susanne F. Yelin. Entanglement devised barren plateau mitigation. *Physical Review Research*, 3:033090, 2021.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5:4213, 2014.

Arthur Pesah, Marco Cerezo, Samson Wang, Tyler Volkoff, Andrew T. Sornborger, and Patrick J. Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11:041011, 2021.

Michael Ragone, Bojko N. Bakalov, Frédéric Sauvage, Alexander F. Kemper, Carlos Ortiz Marrero, Martín Larocca, and M. Cerezo. A lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nature Communications*, 15:7172, 2024.

Stefan H. Sack, Reuben A. Medina, Alexios A. Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding barren plateaus using classical shadows. *PRX Quantum*, 3:020365, 2022.

Frédéric Sauvage, Martín Larocca, Patrick J. Coles, and Marco Cerezo. Building spatial symmetries into parameterized quantum circuits for faster training. *Quantum Science and Technology*, 9: 015029, 2024.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99:032331, 2019.

Andrea Skolik, Jarrod R. McClean, Masoud Mohseni, Patrick van der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence*, 3:5, 2021.

Tyler Volkoff and Patrick J. Coles. Large gradients via correlation in random parameterized quantum circuits. *Quantum Science and Technology*, 6:025008, 2021.

Samson Wang, Enrico Fontana, Marco Cerezo, Krishna Minhazuddin, Akira Sone, Lukasz Cincio, and Patrick J. Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12:6961, 2021.

Roeland Wiersema, Cunlu Zhou, Juan Felipe Carrasquilla, and Yong Baek Kim. Measurement-induced entanglement phase transitions in variational quantum circuits. *SciPost Physics*, 14:147, 2023.

Roeland Wiersema, Efekan Kökcü, Alexander F. Kemper, and Bojko N. Bakalov. Classification of dynamical lie algebras for translation-invariant 2-local spin systems in one dimension. *npj Quantum Information*, 10:110, 2024.

Robert M. Wilcox. Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics*, 8(4):962–982, 1967.

Ge Yan, Hongxu Chen, Kaisen Pan, and Junchi Yan. Rethinking the symmetry-preserving circuits for constrained variational quantum algorithms. In *International Conference on Learning Representations*, 2024.

Kaining Zhang, Liu Liu, Min-Hsiu Hsieh, and Dacheng Tao. Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18612–18627, 2022.

A Exact carrier identities behind Conjecture 1

Conjecture 1 rests on the proved carrier identity, two exact support lemmas, and three ensemble steps that we state but do not prove. The carrier identity is Lemma 1, stated in Section 3.2 and proved below. The one further exact Pauli-algebra statement, the unit support-growth lemma, is proved below. The ensemble steps, a reachability probability for the leading carrier paths, a non-cancellation argument across paths, and a uniform remainder bound fixing σ_0, are collected at the end of this appendix with an account of why each resists the tools we have. This appendix therefore contains the exact layer the conjecture stands on.

Lemma 2 (Unit support growth)**.** *Let Q be a Pauli string and $e^{i\alpha P}$ a two-qubit Pauli evolution on the pair $(q, q+1)$. Then $e^{-i\alpha P} Q e^{i\alpha P}$ is a linear combination of at most two Pauli strings, each with support contained in* $\mathrm{supp}(Q) \cup \{q, q+1\}$*. If* $\mathrm{supp}(Q)$ *is disjoint from $\{q, q+1\}$ the support is unchanged (P and Q commute); otherwise each branch gains at most one site. Consequently, conjugation by one* unit *(four gates on the same pair, by the site lock) grows the support by at most one site, and the carrier's Pauli weight after L layers is at most $O(L)$.*

Proof. For commuting P, Q the conjugation is trivial. For anti-commuting P, Q, $PQ = \pm iR$ with $R = \mp iPQ$ supported on $\{x : P_x Q_x \neq I\}$. Every site of $\mathrm{supp}(Q)$ outside the pair keeps its letter in R, a shared site keeps its letter unless $P_x = Q_x$, and a pair site joins only where Q is absent or $P_x Q_x \neq I$; hence $R \subseteq \mathrm{supp}(Q) \cup \{q, q+1\}$ and $e^{-i\alpha P} Q e^{i\alpha P} = \cos(2\alpha)\, Q \pm \sin(2\alpha)\, R$. No branch gains more than the two pair sites, and a branch can gain sites only when the intersection

is exactly one qubit, in which case at most the second pair site joins. Four gates on the same pair therefore add at most one site per unit, which is where the site lock enters. A unit whose two generators lived on different pairs could add two sites per unit and would degrade the $O(L)$ weight bound to $O(2L)$. □

What is exact. Two elementary facts are used throughout. (i) $\langle 0^n|S|0^n\rangle \neq 0$ only when S is a Z-string, since $|0^n\rangle$ is the joint $+1$ eigenstate of $Z^{\otimes n}$. (ii) In the chain layout every layer's units cover the pair adjacent to the cost site, so the backward-propagated carrier of Lemma 1 meets the cost light cone at every layer.

Why the Napp (2022) estimates do not close the conjecture. The light-cone estimates of Napp (2022) upper-bound the probability that a biased random walk on gate strings reaches a fixed point; they do not lower-bound the ensemble fraction of configurations whose carrier keeps a cost-site component, so the direction of the bound is opposite to what the conjecture needs, and no specialization repairs it. A joint sign reflection of all parameters does not map the ordered product U to $U^\dagger$ either. Negating every parameter sends each unit $R(\theta,\phi) \mapsto R(-\theta,-\phi) = e^{-i\theta A}e^{-i\phi B}e^{i\theta A}e^{i\phi B}$, the inverse of the B-first commutator rather than of R itself, and even unit-wise inversion would not supply the layer reversal that $U^\dagger$ requires, so the reflection parity argument is unavailable and with it a vanishing-mean step. The Gaussian moment $\mathbb{E}[\sin^2(2\phi)] = \frac{1}{2}(1-e^{-8\sigma^2}) = 4\sigma^2 + O(\sigma^4)$ for $\phi \sim \mathcal{N}(0,\sigma^2)$ likewise vanishes as $\sigma \to 0$ and cannot be bounded below by a σ-independent constant, so any assembly must track this factor explicitly.

The open ensemble steps. Closing Conjecture 1 requires three lemmas we do not have. (1) *Path probability.* A polynomial lower bound on the ensemble fraction of configurations in which the carrier retains a cost-site component through L layers, with the polynomial explicit in n and L. (2) *Non-cancellation.* Leading carrier paths carry configuration-dependent signs, and the ensemble variance could in principle cancel across paths; the conjecture asserts a non-cancelling positive fraction, for which orthogonality of the Pauli components is the natural tool. (3) *Uniform remainder and mean control.* A bound on the higher-order BCH corrections uniform in the configuration, holding for $\sigma \leq \sigma_0$ with σ_0 independent of n and L, together with a direct handle on the mean gradient, which can no longer be dismissed by reflection parity.

Status of the conjecture and the distance dependence. What is exact and what is open are as summarized at the start of this appendix. The distance dependence of part (ii) rests on a support count that is exact at the string level. Every Pauli component of the conjugated cost $O' = W^\dagger Z_0 W$ keeps the cost site in its support under the near-identity conjugations of the CoPairSite units, since a conjugation either leaves the cost-site letter untouched or rotates it through the commutator product, which retains the site. A gradient term survives the $|0^n\rangle$ readout only when carrier and cost factors multiply to a Z-string, so the carrier must reach a component of O' whose support equals its own. A carrier at chain distance d from the cost site needs d support rotations on the way in, and each rotation enters the expectation value with one factor of σ^2, the effective order of a unit's BCH generator. The expected gradient at distance d therefore scales as σ^{2+d} times the carrier modulation σ, and the variance as σ^{6+2d}. At $d=0$ this is the conjectured σ^6, at $d=1$ it predicts σ^8, and the gap to a uniform σ^6 bound widens with d. We verified the count by direct Pauli algebra on the nearest motifs; this is why the conjecture is stated with the distance dependence of part (ii).

The admissible range of the conjecture is governed by the initialization scale σ_0. Implicitly, σ_0 is the largest σ for which the order-3 carrier term dominates all order-$\geq$8 corrections in the variance expansion; beyond it the higher-order BCH terms of Eq. (2) begin to compete with the lock. The corrections are generated by products of two or more partner-unit effective generators against the $\sin^2\phi$ branch, so their amplitudes are polynomials in σ of degree ≥ 8 relative to the leading term; the crossover is where σ^2 times the number of competing paths reaches unity. A closed-form bound on σ_0 along these lines is a natural next step. The conjecture concerns the exponent inside the regime, the exact computation of Section 3.3 locates the regime empirically at $\sigma \geq 0.18$, and the measured depth-flat variance at task scales (Section 4) shows the lock itself survives well beyond it.

A short intuition explains why the exponent is 6 rather than 4 or 8. The gradient must read out the Y component of the carrier, which costs one power of the parameters; a single unit cannot rotate Y into a Z-string on its own, since Y has no Z-axis component, so at least one partner unit must act, and

Table 2: Median gradient variance of the commutator-residual ansatz.

cost	σ	n=4	n=6	n=8	n=10	n=12
Z_0	0.1	2.2×10^{-4}	1.3×10^{-4}	1.9×10^{-4}	1.6×10^{-4}	8.5×10^{-5}
Z_0	0.3	8.0×10^{-2}	4.6×10^{-2}	3.4×10^{-2}	2.6×10^{-2}	2.3×10^{-2}
Z_0	0.5	1.5×10^{-1}	6.4×10^{-2}	6.4×10^{-2}	3.0×10^{-2}	2.9×10^{-2}
Z_0	0.7	2.5×10^{-1}	6.6×10^{-2}	3.9×10^{-2}	2.2×10^{-2}	2.4×10^{-2}
TFIM	0.1	3.6×10^{-2}	3.6×10^{-2}	3.7×10^{-2}	3.6×10^{-2}	3.5×10^{-2}
TFIM	0.3	1.5×10^{-1}	1.2×10^{-1}	1.1×10^{-1}	1.1×10^{-1}	9.9×10^{-2}
TFIM	0.5	2.9×10^{-1}	1.5×10^{-1}	9.6×10^{-2}	7.7×10^{-2}	7.2×10^{-2}
TFIM	0.7	5.7×10^{-1}	1.7×10^{-1}	8.1×10^{-2}	4.4×10^{-2}	4.3×10^{-2}

Table 3: Fitted width-scaling exponent $\mathrm{d}\log_{10}(\text{median Var})/\mathrm{d}\log_{10} n$ over $n = 8, \ldots, 16$ at $L = 40$, cost Z_0, with 95% bootstrap intervals from resampling the seeds inside each cell (50 seeds per cell, 100 at $\sigma = 0.1$). Panel D of Figure 2 plots these values. A flat exponent means the variance falls off polynomially in the width, a steep one exponential decay. The HEA exponent at $\sigma = 0.1$ is a crossover value, between its shallow-σ and steep-σ regimes.

	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$	$\sigma = 0.7$
HEA	$-3.14\,[-3.22, -3.08]$	$-7.92\,[-7.96, -7.89]$	$-7.94\,[-7.98, -7.90]$	$-7.91\,[-7.98, -7.89]$
CoPairSite	$-0.59\,[-2.29, 0.67]$	$-0.99\,[-1.43, -0.50]$	$-0.90\,[-1.19, -0.66]$	$-1.11\,[-1.40, -0.79]$

each unit's effective generator is second-order in the parameters; the total gradient order $1 + 2 = 3$ is therefore the shortest path from the locked carrier to the cost diagonal, and the variance inherits its square.

B Numerical evidence for the variance conjectures

Conjecture 1 is itself conjectured at small σ, and Conjecture 2 extends it to task scale, so both rests are measured claims. This appendix supplies that measurement. Table 2 reports the median gradient variance of the commutator-residual ansatz across the Conjecture 2 grid at $L = 40$ (20 initializations per cell; the $L = 12$ task-depth scan shows the same picture), for the local Z_0 cost and the global TFIM cost. The table carries the numerical content of the conjecture. The exponents fitted over the wider $n = 8$ to 16 range are collected in Table 3. At $\sigma \geq 0.3$ the local-Z_0 variance falls off at most polynomially, and under the global TFIM cost it is essentially flat in n at every σ. The order in σ follows the analytic prediction $\mathrm{Var} \sim \sigma^6$ cell by cell.

The small-σ end of Conjecture 1 is tested separately, at $n = 2$, the largest system admitting a decade-wide σ scan (2000 seeds per point). Across $\sigma \in \{0.05, \ldots, 0.18\}$ the ratio $\mathrm{Var}[\partial_\theta C]/\sigma^6$ at $L = 6$ stays in $[454, 631]$, which places the admissible σ_0 at least at 0.18; at fixed $\sigma = 0.12$ the variance *rises* from 1.1×10^{-3} at $L = 2$ to 5.8×10^{-3} at $L = 32$, growth rather than the decay a barren plateau would impose.

C Design choices and ablations

This appendix explains the design choices. The remark separates what is conventional in the construction from what is load-bearing, and the two ablations test the two most natural alternative designs, identity-block initialization and a global cost.

Remark 1 (Sector choice and what is actually load-bearing)**.** *The choice of the Y-odd sector is conventional. Which part of the construction is a gauge convention and which is structural matters for every other sector choice, so we state it precisely. Every single-qubit Clifford relabeling (H swaps $X \leftrightarrow Z$, S maps $X \to Y \to Z \to X$ up to phases) maps the Y-odd pool to an X-odd or Z-odd pool on relabeled qubits, preserving all commutation relations, the CoPairSite anti-commuting-pair structure, the DLA dimension $(4^n - 2^n)/2$, and every subsequent result verbatim. The three single-letter odd gradings are one orbit under the single-qubit Clifford group, and no grading among them is more physical than another. What is structural, not conventional, is threefold. First,* oddness*. The commutator of two odd strings is again odd, which is what makes the DLA a Lie-closed sector with a closed-form dimension; even (majority) pools generate the full algebra with no closed form.*

Second, two-locality on a fixed layout. *The anti-commuting pairs $\{XY, ZY\}$, $\{YX, YZ\}$ exist on every adjacent pair, and their commutators reduce to single-qubit Y, which is what the cross-driving carrier needs to stay single-qubit and experimentally accessible. Third, the* grading–cost interface. *Z-type cost observables (Z_0, TFIM) overlap the carrier sector through exactly the odd-Y rotations analyzed in Appendix A; a purely Z-preserving (codewise even) pool would commute with such costs and have no gradient at first order. No physical symmetry (magnetization parity, particle number) singles out Y-odd over its Clifford orbit; we choose Y for the concrete single-qubit carrier it yields, and a practitioner whose readout basis differs may conjugate the entire construction by the corresponding Clifford at no cost.*

Architecture versus identity-block initialization. HEA admits a near-identity initialization of its own, the Grant-style *identity blocks* (Grant et al., 2021), with all parameters at zero and only the first layer plus a random quarter of the rest perturbed. If the advantage were merely a dressed-up near-identity start, identity-block HEA should match it. We run all three variants under the verdict protocol (cost Z_0, 50 initializations, $n \in \{8, 12\}$, $L \in \{5, 10, 20, 40\}$, $\sigma \in \{0.1, 0.5\}$). Identity-block initialization does repair HEA at shallow depth, reaching 5.6×10^{-3} at $n = 8$, $\sigma = 0.5$, $L = 5$, $2.5\times$ random-init HEA. But the repair decays with depth ($L40/L5$ median ratio 0.32) and width. At $L = 40$ identity-block HEA sits at 1.8×10^{-3} ($n = 8$) and 1.1×10^{-4} ($n = 12$), indistinguishable from random-init HEA (1.9×10^{-3}, 1.2×10^{-4}) and $32\times$ to $287\times$ below CoPairSite (5.6×10^{-2}, 3.3×10^{-2}). Near-identity starts buy shallow-depth relief, not width robustness; the n-scaling contrast of Figure 2 (panel C) is a property of the generator algebra.

Global cost. Local costs are known to soften barren plateaus (Cerezo et al., 2021b); to check that the contrast is not a locality artifact we repeat the scan with the global TFIM Hamiltonian $H = -\sum_i Z_i Z_{i+1} - h \sum_i X_i$ as cost ($n = 8$, 20 initializations, $L \in \{5, 10, 20, 40\}$). The ordering survives with a wider margin. At $\sigma = 0.5$, $L = 40$, CoPairSite's variance is 4.7×10^{-1} against 2.7×10^{-2} for both HEA variants ($18\times$). The global-cost signal, the one optimized in the VQE tasks, is where the margin is largest.

Gradient trajectories through training. The variance comparisons above are ensemble statements at initialization. Figure 3b tracks the same quantity through optimization, the per-step RMS gradient during VQE training at n=12, L=12, σ=0.5, 150 Adam steps, 10 seeds. The seed-averaged RMS gradients are 8.9×10^{-3} for HEA, 4.0×10^{-2} for Naive, and 1.8×10^{-1} for CoPairSite, and the CoPairSite curve remains above HEA's at every step, its worst step still holding a $6\times$ margin. HEA stalls at a mean final cost of -2.07 ± 0.04 while CoPairSite reaches -12.76 ± 0.02 and Naive -12.54 ± 0.10 under the same budget, so the initialization-time gradient advantage is not consumed by training.

D Comparison at controlled budgets

Same-layer-count comparisons conflate architecture with gate budget, and the two gate types are not one hardware resource. A commutator layer costs four two-qubit Pauli evolutions per unit, $4(n - 1)$ per layer, against n CNOTs for an HEA layer, so we match on the parameter count and on the two-qubit-gate count separately rather than on a compiled-equivalent budget. Repeating the gradient-variance scan at comparable budgets ($n = 8$, $K = 400$) confirms the advantage is not a budget artifact under either count. The pairing runs CoPairSite L=6 (84 parameters, 168 two-qubit evolutions) against HEA L=7 (112 parameters, 56 CNOTs), which favors HEA in parameter count and CoPairSite in two-qubit gate count, and the CoPairSite variance is $7\times$ HEA's at $\sigma = 0.3$ (2.5×10^{-2} vs. 3.6×10^{-3}) and $86\times$ at $\sigma = 0.7$ (1.5×10^{-1} vs. 1.8×10^{-3}). Doubling the HEA gate count (HEA L=14 with 112 CNOTs) leaves its variance flat ($\sim 1.9 \times 10^{-3}$ at all σ, so the gap does not track the gate budget). The Naive protocol at the same budget sits between the CoPairSite protocol and HEA, consistent with its partial unit activation. Depth-flatness and absolute signal therefore survive budget-controlled comparison (Figure 6a,b).

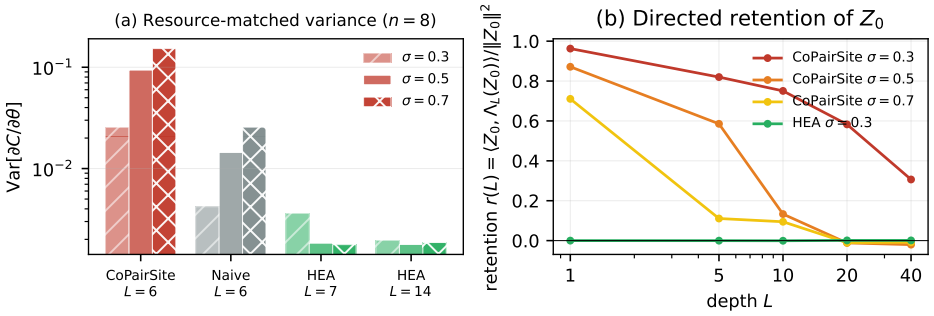


Figure 6: Mechanism checks. (a) Gradient variance under budget-controlled comparison at $n = 8$. The CoPairSite advantage over HEA persists (and grows with σ) at comparable parameter and gate budgets. (b) Directed retention of Z_0 under the averaged circuit channel. The CoPairSite protocol preserves the cost direction sub-exponentially at $\sigma = 0.3$ while HEA destroys it at $L = 1$. Together the two panels show the advantage is not a gate-budget artifact and that the cost direction survives the averaged channel.

Table 4: Lie-closure dimensions by pool design. The one-family pool closes on $4^n/8 + 2^n/4$, a quarter of the full Y-odd sector $(4^n - 2^n)/2$; the order-2 Abelian direction contributes only the n single-qubit Y generators.

	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$
one family $\{XY, ZY\}$	10	36	136	528	2080
other family $\{YX, YZ\}$	10	36	136	528	2080
full Y-odd pool	28	120	496	2016	8128
order-2 only (single Y)	2	3	4	5	6
$4^n/8 + 2^n/4$	10	36	136	528	2080
$(4^n - 2^n)/2$	28	120	496	2016	8128

E CAPACITY, MECHANISM, AND OPERATOR COMPLEXITY

This appendix collects the exact numbers behind Section 4.5. The generator set is fixed by the construction, so the Lie closure is independent of the initialization scale and can be computed once. Table 4 reports the closure dimension for the restricted pools against the full Y-odd sector, computed by exact commutator closure for $n = 3$ to 7.

The order ladder comes from expanding the logarithm of one unit in the shared scale t. For both anti-commuting families the only nonvanishing coefficients at leading order are a single-qubit Y (fitted power 1.98 in t, coefficient $\pm 2i$), with the two pool strings appearing at fitted power 2.98. A one-family pool therefore reaches no entangling generator until third order.

The tangent-space statistics use the exact Jacobian of the circuit map, built from analytic gate derivatives at $n = 5$, $L = 3$ and averaged over three seeds. Table 5 reports the largest singular value divided by σ and the cosine distance between the normalized spectra at $\sigma = 10^{-4}$ and $\sigma = 0.5$. The CoPairSite ratio is constant to four significant figures for $\sigma \leq 10^{-2}$, and its spectrum shape moves by 0.012 across the whole scan, whereas HEA's ratio falls from 1.9×10^4 to 3.1 over the same range.

Training activity is recorded during the TFIM VQE runs at $n = 8$, $L = 12$, $\sigma = 0.5$, 150 Adam steps, 10 seeds, with the activity $|\sin\theta \sin\phi|$ of each unit evaluated at every fifth step. Medians over seeds move from 0.134 to 0.126 for CoPairSite and from 0.371 to 0.495 for HEA; the no-echo control moves from 0.127 to 0.204. The saturated-unit fraction, the share of units with activity above 0.9, moves from 0.03 to 0.06 for CoPairSite against 0.29 to 0.46 for HEA.

The operator-complexity scan decomposes $U^\dagger Z_0 U$ into Pauli strings at $n = 6$, $\sigma = 0.5$. The CoPairSite operator carries no Y-odd component at any depth, and its term count follows the sector size rather than the depth, 116 strings to $L = 6$, 884 to $L = 10$, and 2078 of the 2079 available Y-even strings at $L = 12$. HEA reaches the full 4095 strings by $L = 4$. Operator entanglement, normalized by the half-cut bound $2\min(k, n-k)\ln 2 = 4.159$, rises for CoPairSite from 0.06 at

Table 5: Exact tangent-space spectrum against initialization scale (n=5, L=3, three seeds). Shape distance is the cosine distance of the normalized singular-value vector to its value at $\sigma = 0.5$.

CoPairSite	σ=10^{-4}	10^{-3}	10^{-2}	0.1	0.3	0.5
$\sigma_{\max}/\sigma$	5.786	5.786	5.784	5.654	4.702	3.768
shape distance	0.0120	0.0120	0.0121	0.0099	0.0040	0.0019
HEA						
$\sigma_{\max}/\sigma$	1.9×10^4	1936	194	19.2	5.93	3.12
shape distance	0.0000	0.0000	0.0000	0.0001	0.0006	0.0011

Table 6: DLA dimensions (exact commutator-closure computation). The Y-odd column matches $(4^n - 2^n)/2$ exactly for every computed n.

	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$
HEA ($R_y R_z$ + entangler)	255	1023	4095	16383	65535
unconstrained pool	255	1023	4095	16383	65535
Y-odd pool	120	496	2016	8128	32640
$(4^n - 2^n)/2$	120	496	2016	8128	32640

$L = 4$ to 0.37 at $L = 20$, while HEA is at 0.88 from $L = 6$ onward. The no-echo block sits between them, reaching 0.76 at $L = 20$.

The echo ablation runs $L = 40$ at $\sigma = 0.5$ with 100 seeds at $n = 8$ and 50 seeds at each larger width, and fits the width exponent from the median variance over $n \in \{8, 10, 12, 14, 16\}$. The fitted exponents are -1.46 (CoPairSite), -2.54 (NoEcho), -3.78 (MismatchEcho), and -7.95 (HEA). Only CoPairSite collapses to the identity when a single angle vanishes.

Two further fingerprints separate the matched pairing from a plain four-rotation block, and neither depends on the absolute variance. Fitting the median variance against $\sigma \in \{0.05, \ldots, 0.5\}$ at $n = 6$, $L = 10$ gives a scale exponent of $+5.18$ for CoPairSite, against -0.25 for NoEcho, $+0.62$ for MismatchEcho, and $+1.09$ for HEA; only CoPairSite carries the order-three carrier path, whose variance grows as σ^6. Depth flatness separates them too: the ratio of the median variance at $L = 40$ to its value at $L = 5$, at $n = 8$ and $\sigma = 0.5$, is 0.62 for CoPairSite against 0.13 for NoEcho, 0.10 for MismatchEcho, and 0.89 for HEA. NoEcho is the largest variant at $L = 5$, 3.7×10^{-1} against CoPairSite's 7.7×10^{-2}, but decays by a factor of 7.7 over that depth range and meets CoPairSite by $L = 40$ (4.7×10^{-2} against 4.8×10^{-2}, equal to within seed noise). Its shallow-depth advantage comes from the larger two-qubit pool, not from the echo structure, and it carries neither the σ^6 scale nor the depth flatness. The activity trace of Figure 5b is defined for the echo unit, where $|\sin\theta \sin\phi|$ measures the distance from the identity; a block without the inverse retains an order-one term in θ and ϕ, so the same quantity no longer tracks that distance for NoEcho and its mid-range values are not evidence of protection.

F IDENTITY-UNIT RATES IN FULL PRECISION

This appendix collects the exact values behind the degeneracy analysis of Section 2.2 and the DLA verification of Theorem 1. Table 6 verifies the closed form by exact commutator-closure computation for $n = 4, \ldots, 8$, and the Y-odd column matches $(4^n - 2^n)/2$ in every row. Figure 7 reports the measured probability that a randomly sampled Naive unit is exactly the identity, in full precision, for both pools at $n = 8/10/12/14$; the rate climbs toward unity with width. Figure 8 illustrates the four sampling protocols on a six-qubit layer, the geometric reason Naive wastes the parameter budget (disjoint supports give $[A, B] = 0$ regardless of strings) and how the pair and site locks repair it.

G PROOFS OF THE SUPPORTING LEMMAS

Theorem 1 assembles from supporting lemmas, and the exact layer of Conjecture 1 rests on the proved carrier and support lemmas of Appendix A. This appendix collects the four supporting proofs in the order in which the lemmas appear, together with the parity identity behind the DLA closure.

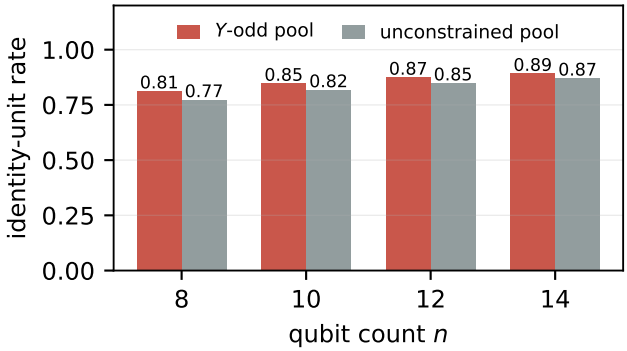


Figure 7: Measured probability that a sampled unit is exactly the identity ($[A, B] = 0$), Naive independent sampling. 2×10^5 samples per entry. The rate climbs toward unity with width for both pools, and it is slightly higher for the Y-odd pool. Full precision, $0.813/0.847/0.875/0.892$ (Y-odd pool) and $0.773/0.818/0.849/0.871$ (unconstrained pool) at $n = 8/10/12/14$.

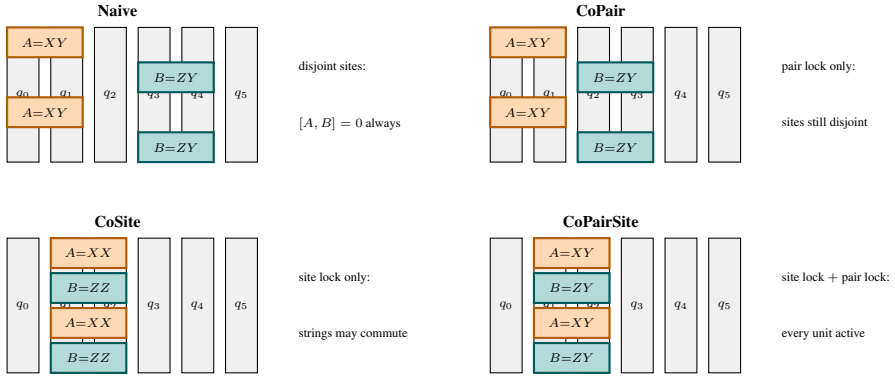


Figure 8: The four sampling protocols of the 2×2 family on a six-qubit layer. Each panel shows one sampled unit, four gates built from generators A (orange) and B (teal), applied in alternating order on their supports. Naive draws sites and strings independently; disjoint supports (A on (q_0, q_1), B on (q_3, q_4)) always commute, so the unit collapses to the identity. CoPair locks A, B to an anti-commuting pair (A=XY, B=ZY) but the sites remain disjoint and still commute. CoSite forces all four gates onto one pair, but independent strings may commute (A=XX, B=ZZ). CoPairSite applies both locks, so every unit is active.

Lemma 3 (Pauli Y-parity identity)**.** *For Pauli strings P, Q let $y(P)$ denote the number of Y factors of P. Then*

$$y(PQ) \;\equiv\; y(P) + y(Q) + \mathbf{1}[P, Q \text{ anticommute}] \pmod 2. \tag{8}$$

In particular, if P and Q are both Y-odd and anticommute, then PQ is Y-odd.

Proof of Lemma 3. Let n_{ab} count the qubits on which P carries a and Q carries b, $a, b \in \{I, X, Y, Z\}$. Positionwise multiplication gives $PQ = \omega \bigotimes_q T_q$ with $\omega \in \{\pm 1, \pm i\}$ and

$$y(PQ) \;=\; n_{XZ} + n_{ZX} + n_{YI} + n_{IY}, \tag{9}$$

since only the pairs (X, Z), (Z, X), (Y, I), (I, Y) produce a Y factor. The anticommutation condition reads $\mathbf{1}[\text{anti}] = \big(n_{XZ} + n_{ZX} + n_{XY} + n_{YX} + n_{YZ} + n_{ZY}\big) \bmod 2$, and $y(P) = n_{YX} + n_{YY} + n_{YZ} + n_{YI}$, $y(Q) = n_{XY} + n_{YY} + n_{ZY} + n_{IY}$. Summing,

$$\begin{aligned} y(P) + y(Q) + \mathbf{1}[\text{anti}] &\equiv n_{XZ} + n_{ZX} + n_{YI} + n_{IY} + 2\big(n_{XY} + n_{YX} + n_{YZ} + n_{ZY} + n_{YY}\big) \\ &\equiv y(PQ) \pmod 2, \end{aligned} \tag{10}$$

which is Equation (8). If P and Q are both Y-odd ($y(P) \equiv y(Q) \equiv 1$) and anticommute, then $y(PQ) \equiv 1 + 1 + 1 \equiv 1$. The product is Y-odd, and so is the Pauli type of $[P, Q] = PQ - QP = 2PQ$. □

Lemma 4 (Gradient duality)**.** *For $C = \langle 0|U^\dagger O U|0\rangle$ with $O = Z_0$,*

$$\frac{\partial C}{\partial \theta} = 2\,\mathrm{Im}\,\langle 0|\,V^\dagger\big(e^{-i\phi B} A e^{i\phi B} - A\big)\,O'\,V|0\rangle, \tag{11}$$

$$\frac{\partial C}{\partial \phi} = -2\,\mathrm{Im}\,\langle 0|\,G_4^\dagger\,\big[B,\,e^{i\theta A}\big]\,O'\,V|0\rangle. \tag{12}$$

Each driving term contains only the other *parameter, an exact duality.*

Proof of Lemma 4. With $|\psi\rangle = U|0\rangle = WV|0\rangle$, $\partial_\theta C = 2\,\mathrm{Re}\,\langle \partial_\theta\psi|O|\psi\rangle$ and $\partial_\theta\psi = iA\,U|0\rangle - i\,WAV|0\rangle$, using $\partial_\theta G_1 = iAG_1$ and $\partial_\theta G_3 = -iAG_3$ (the generator commutes with its own evolution). Hence

$$\begin{aligned}\langle \partial_\theta\psi|O|\psi\rangle &= -i\langle\psi|AO|\psi\rangle + i\langle 0|V^\dagger A\,O'\,V|0\rangle \\ &= -i\langle 0|V^\dagger\big(W^\dagger AW - A\big)O'V|0\rangle = -i\langle 0|V^\dagger\big(G_2^\dagger AG_2 - A\big)O'V|0\rangle,\end{aligned} \tag{13}$$

where $W^\dagger AOW = (W^\dagger AW)(W^\dagger OW)$ and $W^\dagger AW = G_2^\dagger AG_2$ because A commutes with $G_1 = e^{i\theta A}$. Taking $2\,\mathrm{Re}(-i\,z) = 2\,\mathrm{Im}\,z$ yields Eq. (11). For Eq. (12), $\partial_\phi\psi = iWB\,V|0\rangle - iWG_3B\,G_4|0\rangle$ (using $BG_2 = G_2B$), so

$$\begin{aligned}\langle \partial_\phi\psi|O|\psi\rangle &= -i\langle 0|V^\dagger B\,O'V|0\rangle + i\langle 0|G_4^\dagger B\,G_3^\dagger O'V|0\rangle \\ &= i\langle 0|G_4^\dagger\big(BG_3^\dagger - G_3^\dagger B\big)O'V|0\rangle = i\langle 0|G_4^\dagger\big[B,\,e^{i\theta A}\big]O'V|0\rangle,\end{aligned} \tag{14}$$

and $2\,\mathrm{Re}(i\,z) = -2\,\mathrm{Im}\,z$ yields Eq. (12). □

Proof of Lemma 1. For anti-commuting A, B ($\{A, B\} = 0$), $[A, B] = 2AB$, and the rotation identity $e^{-i\phi B}Ae^{i\phi B} = \cos(2\phi)A + i\sin(2\phi)AB$ follows from expanding the exponential and using $[A, B] = 2AB$ termwise, or directly from the $SO(3)$ action of B on A. For the CoPairSite pairs the two-qubit product AB equals $\pm iY$, where Y is the single-qubit Pauli-Y on the site at which A and B disagree, because the pair coincides on the shared-Y site (contributing $Y^2 = I$), and on the disagreeing site contributes $XZ = \mp iY$ (or $ZX = \pm iY$). Hence $i\sin(2\phi)AB = \pm\sin(2\phi)Y$, giving (4). Substituting into (11) and separating the $\sin^2$ and $\sin(2\phi)$ terms yields (5). For (6), $e^{i\theta A} = \cos\theta + i\sin\theta\,A$ exactly (since $A^2 = I$), so $[B, e^{i\theta A}] = i\sin\theta\,[B, A] = -2i\sin\theta\,AB = \pm 2\sin\theta\,Y$; substituting into (12) gives the result. □

Lemma 5 (Four-point parameter-shift rule)**.** *Fix all circuit parameters except one unit parameter θ. The cost $C(\theta)$ is a trigonometric polynomial with frequency support $\{0, \pm 2, \pm 4\}$, and*

$$\frac{\partial C}{\partial \theta} = \big(1 - \sqrt{2}\big)\,\big[C\big(\theta + \tfrac{\pi}{4}\big) - C\big(\theta - \tfrac{\pi}{4}\big)\big] + 2\,\big[C\big(\theta + \tfrac{\pi}{8}\big) - C\big(\theta - \tfrac{\pi}{8}\big)\big]. \tag{15}$$

Four circuit evaluations per parameter, fixed shifts and weights, no system solve, exact in exact arithmetic.

Proof of Lemma 5. Frequency support. Fix all parameters except one unit parameter θ, and collect all θ-independent factors. The ket side contains the two gates $e^{i\theta A}$ and $e^{-i\theta A}$ of the unit; the bra side ($U^\dagger$) contains the conjugate pair $e^{i\theta A}e^{-i\theta A}$ in reverse order. Each gate $e^{is\theta A}$ with $s = \pm 1$ has matrix elements $\cos(s\theta)\langle u|v\rangle + i\sin(s\theta)\langle u|Av\rangle$ for eigenvectors $|v\rangle$, $Av\rangle$ of A ($A^2 = I$), i.e., it contributes a first-order trigonometric factor in θ. With four such factors, $C(\theta)$ is a linear combination of products $\prod_j \mathrm{trig}(s_j\theta)$ with $s_j \in \{\pm 1\}$, so every term has frequency in $\{0, 2, 4\}$. Writing

$$C(\theta) = c_0 + c_2\cos 2\theta + s_2\sin 2\theta + c_4\cos 4\theta + s_4\sin 4\theta, \tag{16}$$

the derivative components are $X_2 = -c_2\sin 2\theta + s_2\cos 2\theta$, $X_4 = -c_4\sin 4\theta + s_4\cos 4\theta$, and $\partial_\theta C = 2X_2 + 4X_4$.

Shifted differences. Using $\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta$ and the sine counterpart,

$$D(s) \equiv C(\theta + s) - C(\theta - s) = 2\sin(2s)\,X_2 + 2\sin(4s)\,X_4. \tag{17}$$

At $s = \pi/4$, $\sin(\pi/2) = 1$ and $\sin(\pi) = 0$, so $X_2 = D(\pi/4)/2$. At $s = \pi/8$, $\sin(\pi/4) = \sqrt{2}/2$ and $\sin(\pi/2) = 1$, so $X_4 = D(\pi/8)/2 - (\sqrt{2}/2)\,X_2$. Substituting,

$$\partial_\theta C = 2X_2 + 4X_4 = D(\pi/4) + 2D(\pi/8) - \sqrt{2}\,D(\pi/4) = \big(1 - \sqrt{2}\big)D\big(\tfrac{\pi}{4}\big) + 2D\big(\tfrac{\pi}{8}\big), \tag{18}$$

which is Eq. (15).

Machine verification. We verified Eq. (15) on randomized deep circuits, with $n \in \{4, 5, 6\}$, $L \in \{2, \ldots, 6\}$, both the CoPairSite and unconstrained pools, both unit parameters, single-qubit observables $\{X, Y, Z\}$ on random qubits, and initialization scales $\sigma \in \{0.1, 0.3, 0.5, 0.8, 3.0\}$. Ground truth is spectral differentiation of a dense 256-point sample of $C(\theta)$ (exact up to round-off, independent of the rule under test). Over 500 randomized trials the maximum absolute deviation of Eq. (15) from the spectral derivative is 1.4×10^{-13}; against central finite differences ($h = 10^{-7}$) it is 7.1×10^{-9} at amplitudes where the finite-difference truncation error itself dominates. The verification script accompanies the supplementary code. □

Remark (why not the standard two-point rule). The standard rule $\partial_\theta C = \frac{1}{2}[C(\theta + \frac{\pi}{2}) - C(\theta - \frac{\pi}{2})]$ applies when C depends on θ only through a single generator $e^{-i\theta G}$ with eigenvalues ± 1 of G, so that C carries frequency ± 1. Here θ appears in four gates with mixed signs, producing frequencies $\{2, 4\}$; at shift $\pi/2$ the two-point difference vanishes identically, since $\sin(2s) = \sin \pi = 0$ and $\sin(4s) = \sin 2\pi = 0$, and no single shift separates the two frequency components. Lemma 5 restores exactness at the cost of two additional evaluations.

H Completeness of the generator set

Remark. $\dim \mathfrak{g}_{\mathrm{y}} = (4^n - 2^n)/2 = 2^{n-1}(2^n - 1)$ coincides with $\dim \mathfrak{so}(2^n)$; no isomorphism claim is made or needed.

We prove that the adjacent-pair Y-odd pool generates every Y-odd Pauli string, i.e. $\mathfrak{g}_{\mathrm{y}} \subseteq \mathfrak{g}_{\mathrm{y}}$, which together with the closure direction of Theorem 1 gives equality. Encode a Pauli string by $(a, b) \in \mathbb{F}_2^{2n}$ with $X = (1, 0)$, $Z = (0, 1)$, $Y = (1, 1)$, $I = (0, 0)$. Then the Y-parity is $|P|_Y = a \cdot b \pmod 2$, and two strings anticommute iff $\omega(P, Q) := a \cdot b_Q + b \cdot a_Q = 1 \pmod 2$; in that case the commutator type is the pointwise product $PQ = (a + a_Q, b + b_Q)$.

Lemma 6 (Same-position commutators generate single-qubit Y)**.** *For each $1 \le i \le n$ the generator set contains (up to phase) the single-qubit Pauli Y_i.*

Proof. Same-position anti-commuting pairs give $[X_i Y_{i+1}, Z_i Y_{i+1}] \propto -Y_i$ and $[Y_i X_{i+1}, Y_i Z_{i+1}] \propto -Y_{i+1}$, since the shared Y_{i+1} (resp. Y_i) factor cancels in the commutator because $Y \cdot Y = I$. The endpoint qubits follow from the two families combined ($i = 1$ from the first family, $i = n$ from the second). □

Base case $n = 2$. The six Y-odd strings on two qubits are the four generators $\{XY, YX, YZ, ZY\}$ and the two singles $\{YI, IY\}$ of Lemma 6, so $\mathfrak{g}_2 = \mathfrak{g}_{\mathrm{y}}$, $\dim = 6$.

Induction step $n - 1 \to n$. Assume $\mathfrak{g}_{n-1} = \mathfrak{g}_{\mathrm{y}}|_{n-1}$. Since the pool on qubits $1..n-1$ is a sub-pool, $\mathfrak{g}_{n-1} \subseteq \mathfrak{g}_n$. Take any Y-odd string S on n qubits and write $S = Q' \otimes T$ with Q' on qubits $1..n-3$ and $T \in \{I, X, Y, Z\}^3$ on the window $(n-2, n-1, n)$.

If T ends in I, then S is supported on $1..n-1$, so $S \in \mathfrak{g}_{n-1} \subseteq \mathfrak{g}_n$. Otherwise T is one of the 48 window patterns ending in a non-identity, and Table 7 gives a start pattern T_0 (ending in I, same Y-parity as T) together with a bracket sequence $T_0 = T^{(0)} \to T^{(1)} \to \cdots \to T^{(k)} = T$, $k \le 3$, where each step is $T^{(r+1)} = [T^{(r)}, P_r]$ and P_r is one of the 11 allowed window partners, namely the 8 pool generators on edges $(n-2, n-1)$ and $(n-1, n)$, and the singles Y_{n-2}, Y_{n-1}, Y_n. Each partner lies in $\mathfrak{g}_n$ (the generators by definition, Y_{n-2}, Y_{n-1} by induction, Y_n by Lemma 6). Since $\mathrm{par}(T_0) = \mathrm{par}(T)$ and S is Y-odd, $S_0 := Q' \otimes T_0$ is Y-odd and supported on $1..n-1$, so $S_0 \in \mathfrak{g}_{n-1} \subseteq \mathfrak{g}_n$. The partners act only inside the window, so Q' is preserved, and replaying the sequence $S_0 \xrightarrow{[\cdot, P_0]} S_1 \to \cdots \to S_k = Q' \otimes T = S$ shows $S \in \mathfrak{g}_n$.

The induction is valid for all $n \ge 3$ from the base $n = 2$; the window $(n-2, n-1, n)$ is a sub-chain, so no quantity depends on n. The finite table is the only non-elementary ingredient and is verified exhaustively by an independent script, which replays every bracket (each is checked non-vanishing and the final product equals the target).

Table 7: The complete 48-entry three-qubit attach table, printed in full. Each target tail T (ending in a non-identity) is reached from a start T_0 (ending in I, same Y-parity) by at most 3 commutators with the pool generators on edges $(n-2, n-1)$, $(n-1, n)$ and the singles Y_{n-2}, Y_{n-1}, Y_n (window notation, abc on $(n-2, n-1, n)$). Every row is replayed and verified by a script in the supplementary code. The depth distribution is $28/15/5$ at depths $1/2/3$.

T	T_0	k	sequence	T	T_0	k	sequence
$I\,I\,X$	$I\,X\,I$	2	IYZ, IZY	$Y\,I\,X$	$Y\,X\,I$	2	IYZ, IZY
$I\,I\,Y$	$Y\,X\,I$	3	IYX, YXI, IYZ	$Y\,I\,Y$	$X\,X\,I$	2	IZY, ZYI
$I\,I\,Z$	$I\,X\,I$	2	IYX, IZY	$Y\,I\,Z$	$Y\,X\,I$	2	IYX, IZY
$I\,X\,X$	$I\,Z\,I$	1	IYX	$Y\,X\,X$	$Y\,Z\,I$	1	IYX
$I\,X\,Y$	$I\,Y\,I$	1	IZY	$Y\,X\,Y$	$Y\,Y\,I$	1	IZY
$I\,X\,Z$	$I\,Z\,I$	1	IYZ	$Y\,X\,Z$	$Y\,Z\,I$	1	IYZ
$I\,Y\,X$	$Y\,X\,I$	2	IYX, YXI	$Y\,Y\,X$	$I\,X\,I$	2	IYX, YXI
$I\,Y\,Y$	$I\,X\,I$	1	IZY	$Y\,Y\,Y$	$Y\,X\,I$	1	IZY
$I\,Y\,Z$	$Y\,X\,I$	2	IYZ, YXI	$Y\,Y\,Z$	$I\,X\,I$	2	IYZ, YXI
$I\,Z\,X$	$I\,X\,I$	1	IYX	$Y\,Z\,X$	$Y\,X\,I$	1	IYX
$I\,Z\,Y$	$I\,Y\,I$	1	IXY	$Y\,Z\,Y$	$Y\,Y\,I$	1	IXY
$I\,Z\,Z$	$I\,X\,I$	1	IYZ	$Y\,Z\,Z$	$Y\,X\,I$	1	IYZ
$X\,I\,X$	$X\,X\,I$	2	IYZ, IZY	$Z\,I\,X$	$X\,X\,I$	2	IYX, YZI
$X\,I\,Y$	$Y\,X\,I$	2	IZY, ZYI	$Z\,I\,Y$	$X\,Y\,I$	2	IXY, YZI
$X\,I\,Z$	$X\,X\,I$	2	IYX, IZY	$Z\,I\,Z$	$X\,X\,I$	2	IYZ, YZI
$X\,X\,X$	$X\,Z\,I$	1	IYX	$Z\,X\,X$	$Z\,Z\,I$	1	IYX
$X\,X\,Y$	$X\,Y\,I$	1	IZY	$Z\,X\,Y$	$Z\,Y\,I$	1	IZY
$X\,X\,Z$	$X\,Z\,I$	1	IYZ	$Z\,X\,Z$	$Z\,Z\,I$	1	IYZ
$X\,Y\,X$	$Y\,X\,I$	3	IYZ, IZY, ZYI	$Z\,Y\,X$	$X\,Y\,I$	3	IXY, YZI, IYZ
$X\,Y\,Y$	$X\,X\,I$	1	IZY	$Z\,Y\,Y$	$Z\,X\,I$	1	IZY
$X\,Y\,Z$	$Y\,X\,I$	3	IYX, IZY, ZYI	$Z\,Y\,Z$	$X\,Y\,I$	3	IXY, YZI, IYX
$X\,Z\,X$	$X\,X\,I$	1	IYX	$Z\,Z\,X$	$Z\,X\,I$	1	IYX
$X\,Z\,Y$	$X\,Y\,I$	1	IXY	$Z\,Z\,Y$	$Z\,Y\,I$	1	IXY
$X\,Z\,Z$	$X\,X\,I$	1	IYZ	$Z\,Z\,Z$	$Z\,X\,I$	1	IYZ

A worked example. Take the target tail $T = ZIZ$ from Table 7 (start $T_0 = XXI$, sequence IYZ then YZI). In window notation on three qubits, with phases discarded,

$$XXI \xrightarrow{[\,\cdot\,, IYZ]} XZZ, \qquad XZZ \xrightarrow{[\,\cdot\,, YZI]} ZIZ.$$

With phases kept, each bracket equals $2i$ times the Pauli product of its two strings. The first step is $[XX, IY] \otimes Z = 2i\,(XZ) \otimes Z = 2i\,XZZ$, and the second is $[XZ, YZ] \otimes I = 2i\,ZI \otimes I$, whose type matches ZIZ after the shared window factors are read off. Each partner anticommutes with the current string on exactly one window factor, so no step vanishes. Each partner is either a pool generator (IYZ and YZI live on the window edges) or, in the general table, an inductively available single-qubit Y. The start XXI ends in I, so $S_0 = Q' \otimes XXI$ is supported on qubits $1..n-1$ whenever Q' is, and the replay never touches Q'.

Remark. Because $Y_j \in \mathfrak{g}_n$ for every j, conjugation by $e^{i\theta Y_j}$ swaps $X \leftrightarrow Z$ on qubit j, so $X \leftrightarrow Z$ on any qubit is a symmetry of $\mathfrak{g}_y$ and could compress Table 7; we state it uncompressed for clarity.

I Experimental details

All simulations use a pure state-vector simulator with adjoint-state gradients and layerwise checkpointing, validated against central differences to $< 10^{-9}$ absolute error on all tested configurations. DLA dimensions are computed by exact commutator-closure breadth-first search over Pauli strings with phase discarded. Unit-degeneracy rates use 2×10^5 Monte Carlo samples per configuration. The verdict experiments use $n \in \{8, 10, 12, 14, 16\}$, $L \in \{5, 10, 20, 40\}$, $\sigma \in \{0.1, 0.5\}$, 50 random initializations, cost Z_0; 3600 runs at $n \leq 12$ plus 2400 runs at $n \in \{14, 16\}$.

Statistical reporting. Gradient-variance ensembles in these experiments are heavy-tailed. Naive and CoSite circuits can be *structurally zero* (every sampled unit an identity; variance below 10^{-30}), and live seeds span several orders of magnitude. We therefore report, for the verdict experiment, the median over nonzero seeds and the structural zero fraction in Table 8; the corresponding means

Table 8: Verdict experiment at $n = 12$, $\sigma = 0.5$, showing the median gradient variance over nonzero seeds (structural-zero fraction in parentheses, i.e., circuits in which every unit is an identity) and depth-retention ratio $L{=}40/L{=}5$. Medians throughout, since Naive's structural zeros make its mean unstable. The two single-lock variants bracket the full protocol from below, isolating each lock dimension.

variant	$L = 5$	$L = 10$	$L = 20$	$L = 40$	$L40/L5$
Naive Y-odd	$1.2 \times 10^{-3}\,(74\%)$	$5.0 \times 10^{-3}\,(30\%)$	$4.8 \times 10^{-3}\,(4\%)$	$3.9 \times 10^{-3}\,(0\%)$	3.3
CoPair	$4.5 \times 10^{-3}\,(38\%)$	$5.7 \times 10^{-3}\,(10\%)$	$7.6 \times 10^{-3}\,(2\%)$	$8.4 \times 10^{-3}\,(0\%)$	1.9
CoSite	$7.4 \times 10^{-3}\,(30\%)$	$2.1 \times 10^{-2}\,(2\%)$	$2.2 \times 10^{-2}\,(0\%)$	$1.9 \times 10^{-2}\,(0\%)$	2.6
CoPairSite	$4.4 \times 10^{-2}\,(0\%)$	$4.9 \times 10^{-2}\,(0\%)$	$4.5 \times 10^{-2}\,(0\%)$	$2.7 \times 10^{-2}\,(0\%)$	0.61
HEA ($\sigma = 0.5$)	exponential decay with L				$\ll 1$

Table 9: TFIM phase-recognition test accuracy (mean$\pm$s.d. over 10 seeds, 100 test states, $n = 8$, $\sigma = 0.5$). The CoPairSite protocol dominates at every depth, and its margin over the Naive protocol widens with depth as the latter saturates.

	$L = 8$	$L = 12$	$L = 16$
HEA	0.860 ± 0.037	0.891 ± 0.033	0.888 ± 0.027
Naive Y-odd	0.721 ± 0.069	0.767 ± 0.082	0.816 ± 0.103
CoPairSite	$\mathbf{0.900} \pm 0.074$	$\mathbf{0.957} \pm 0.025$	$\mathbf{0.971} \pm 0.025$

are given in Table 11 and agree with the medians on all qualitative conclusions (orderings between variants and depth-flatness).

Two gradient metrics. The quantity reported throughout, $\frac{1}{P}\sum_j g_j^2$ averaged over initializations, is a *mean-squared-gradient signal*. The strict barren-plateau quantity is the per-parameter variance across initializations, $\frac{1}{P}\sum_j \mathrm{Var}(g_j)$. We compared both metrics at $n \in \{8, 12\}$, $L \in \{5, 10, 20, 40\}$, $\sigma \in \{0.1, 0.5\}$, 50 initializations. The two agree to within a per-parameter bias fraction $\big(\mathbb{E}[g_j]\big)^2/\mathbb{E}[g_j^2]$ of at most 3% in every configuration, so the per-seed gradient distribution is centered and the signal metric and the strict variance coincide in ordering and in magnitude; all conclusions are metric-independent. At $L = 40$, $\sigma = 0.5$, the CoPairSite-to-HEA ratio is $277\times$ under the signal metric and $278\times$ under the strict variance metric at $n = 12$.

This appendix has listed the full-precision values behind the experiments of Section 4, namely the protocol-verdict medians of Section 4.1 (Table 8), the VQE energies of Section 4.2 (Table 10), the classification accuracies of Section 4.3 (Table 9), and the noise scan cited in Section 4.4 (Table 12 and Figure 9).

Table 10: VQE final energies (mean±s.d. over 10 seeds, $n = 8$, $L = 12$, 150 Adam steps). Ground-state energies are $E_0 = -8.509$ (TFIM) and $E_0 = -14.604$ (Heisenberg). On TFIM the CoPairSite ansatz converges at every initialization scale where HEA collapses (the HEA failure at $\sigma \geq 0.5$ is consistent across all seeds); on Heisenberg the CoPairSite ansatz is closest to E_0 at every σ and improves monotonically with initialization scale.

	TFIM			Heisenberg		
variant	σ=0.3	0.5	0.7	σ=0.3	0.5	0.7
HEA	$-8.503\pm.002$	-7.36 ± 0.79	-6.85 ± 0.08	-12.47 ± 1.21	-12.03 ± 0.96	-11.57 ± 0.14
Naive	$-8.431\pm.038$	$-8.441\pm.050$	$-8.455\pm.040$	-12.78 ± 0.28	-12.69 ± 0.42	-12.64 ± 0.44
CoPairSite	$\mathbf{-8.505}\pm.005$	$\mathbf{-8.505}\pm.003$	$\mathbf{-8.501}\pm.009$	$\mathbf{-13.23}\pm0.10$	$\mathbf{-13.31}\pm0.05$	$\mathbf{-13.49}\pm0.22$

Table 11: Mean gradient variance over all seeds, structural zeros included, for the verdict experiment at $n = 12$, the mean-metric counterpart of the medians in Table 8. CoPairSite remains the top variant at every depth and both initialization scales, and no variant decays exponentially with depth.

	$L = 5$	$L = 10$	$L = 20$	$L = 40$
$\sigma = 0.5$				
Naive Y-odd	3.3×10^{-3}	8.5×10^{-3}	6.0×10^{-3}	4.6×10^{-3}
CoPair	1.1×10^{-2}	9.5×10^{-3}	1.0×10^{-2}	9.0×10^{-3}
CoSite	1.2×10^{-2}	2.5×10^{-2}	2.2×10^{-2}	2.1×10^{-2}
CoPairSite	6.2×10^{-2}	5.8×10^{-2}	4.4×10^{-2}	3.3×10^{-2}
$\sigma = 0.1$				
Naive Y-odd	1.8×10^{-7}	6.8×10^{-6}	5.8×10^{-6}	7.5×10^{-6}
CoPair	8.5×10^{-6}	4.7×10^{-6}	1.9×10^{-5}	2.4×10^{-5}
CoSite	1.9×10^{-6}	3.0×10^{-5}	2.1×10^{-5}	3.6×10^{-5}
CoPairSite	1.1×10^{-4}	8.1×10^{-5}	1.1×10^{-4}	2.7×10^{-4}

Table 12: Mean gradient variance at $L = 40$, $n = 8$, under local depolarizing noise with per-qubit error probability p ($K = 20$ trajectories $\times$ 20 initializations). Full-precision values underlying Section 4.4.

	$p = 0$	$p = 0.002$	$p = 0.005$	$p = 0.01$
$\sigma = 0.5$				
HEA	1.9×10^{-3}	6.4×10^{-4}	1.8×10^{-4}	9.5×10^{-5}
Naive Y-odd	1.4×10^{-2}	1.0×10^{-2}	5.8×10^{-3}	3.0×10^{-3}
CoPairSite	4.3×10^{-2}	3.1×10^{-2}	1.7×10^{-2}	9.6×10^{-3}
$\sigma = 0.1$				
HEA	2.8×10^{-3}	1.3×10^{-3}	5.1×10^{-4}	2.1×10^{-4}
Naive Y-odd	4.3×10^{-5}	3.6×10^{-5}	2.7×10^{-5}	1.1×10^{-5}
CoPairSite	4.5×10^{-4}	4.1×10^{-4}	2.5×10^{-4}	1.8×10^{-4}

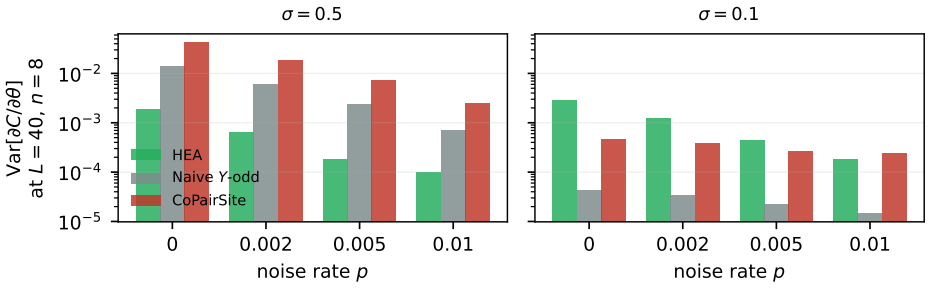


Figure 9: Mean gradient variance at L=40, n=8, under the local depolarizing channel, against the per-qubit error probability p (K=20 trajectories $\times$ 20 initializations, log scale), the figure form of Table 12. At the working initialization σ=0.5 (left) CoPairSite exceeds HEA by 22–101$\times$ across the noise range. At σ=0.1 (right), below the admissible scale $\sigma_0 \geq 0.18$ of Conjecture 1, both architectures sit in the narrow-initialization regime and the two approach each other as p grows.