Journal of High School Science

Gradient usability in few-qubit quantum Neural Networks: A signal-to-noise ratio framework for evaluating mitigation strategies

Shen B[1], Ling K[2]

Submitted: March 26, 2026, Revised: version 1, May 29, 2026, version 2, July 26, 2026
Accepted: July 27, 2026

Abstract
The barren-plateau problem—the exponential flattening of optimization landscapes—severely limits the trainability of quantum neural networks (QNNs) on near-term hardware. This review examines ansatz-design and complementary mitigation strategies for few-qubit QNNs under finite-shot noisy intermediate-scale quantum (NISQ) constraints. Its primary conceptual contribution is a literature-grounded reframing of gradient-based trainability around pointwise finite-shot gradient signal-to-noise ratio (SNR) while retaining exact-gradient landscape variance as a complementary theoretical diagnostic. This reframing follows because gradient-based optimization acts on finite-shot gradient estimates whose statistical resolvability directly determines whether the optimizer receives a reliably distinguishable update signal, whereas numerical studies restricted to 2–10 qubits generally cannot establish asymptotic exponential or polynomial gradient-decay laws without theoretical support. Estimator SNR therefore addresses the immediate experimental question of whether the update signal is distinguishable from its sampling uncertainty, whereas exact-gradient variance characterizes landscape concentration. The review surveys ansatz design, parameter initialization, cost-function selection, and error management, and develops a conditional design heuristic: problem-inspired ansätze are preferred when the task maps onto circuit families with demonstrated trainability properties; otherwise, shallow hardware-efficient ansätze provide controlled baselines. As a secondary contribution, the review proposes a complete eight-configuration 2^3 factorial framework for evaluating interactions among restricted entanglement, local cost functions, and active residual shortcuts under matched measurement budgets. An exact identity motivates testing the interaction between restricted entanglement and local costs at the exact-gradient-signal level, but the interaction is tested two-sided; all estimator-SNR interactions remain open empirical questions. The framework is presented as an implementable research agenda enabled by the SNR perspective rather than as evidence that any interaction has been established.

Keywords
Quantum neural networks, Quantum machine learning, Variational quantum algorithms, Barren plateaus, Gradient signal-to-noise ratio, Trainability, NISQ computing, Variational quantum circuits, Quantum optimization, Ansatz design

[1]Corresponding author: Brandon Shen, Carlmont High School, 1400 Alameda de las Pulgas, Belmont, CA 94002, USA. brandonshen123@gmail.com
[2]Karena Ling, Carlmont High School, 1400 Alameda de las Pulgas, Belmont, CA 94002, USA. karenaxling@gmail.com

https://doi.org/10.64336/001c.166201

1. Introduction

1.1 Background & applications of QNNs

The development of variational quantum algorithms, including quantum neural networks (QNNs), was motivated by the practical limitations of near-term quantum hardware. Peruzzo et al. (1) introduced the Variational Quantum Eigensolver (VQE), a hybrid quantum–classical model in which a quantum processor prepares and measures a parameterized state—an ansatz—while a classical optimizer updates its parameters through a feedback loop. By treating circuit parameters as learnable variables, this variational structure provided a foundation for QNNs and related machine-learning applications. Unlike classical bits, qubits can occupy superposition states; consequently, an n-qubit state is represented in a Hilbert space whose dimension grows as 2^n.

The primary motivation for developing QNNs and other quantum machine-learning (QML) methods is their potential to enhance data analysis and address problems beyond the practical reach of classical approaches. Under specified assumptions, some QML methods have been proposed to perform linear-algebraic operations with resources that scale logarithmically in the number of features or training examples (2). Quantum methods have also been proposed for optimization and generative modeling, including approaches that map learning problems onto low-energy configurations of physical systems (2).

This review assumes familiarity with basic linear algebra and probability but no prior background in quantum computing. Key concepts—including qubits, superposition, and circuit structure—are introduced as they become relevant. The central challenge is an optimization problem: how to design a learning system whose training signal remains informative long enough for meaningful learning to occur.

Hybrid quantum–classical optimization also introduces a critical limitation: gradient vanishing. As circuit depth or qubit count increases, the cost landscape of a variational quantum algorithm can flatten exponentially, causing gradients to vanish and rendering optimization ineffective (3). This phenomenon, known as a barren plateau, resembles vanishing-gradient behavior in classical deep networks but arises through quantum concentration mechanisms associated with rapidly expanding state spaces and circuit structure.

The exponential growth of Hilbert-space dimension is central both to the representational power of quantum models and to the concentration phenomena associated with barren plateaus (3). Noisy intermediate-scale quantum (NISQ) devices add finite sampling, decoherence, and connectivity constraints to this optimization problem. Consequently, despite their theoretical promise, QNNs face substantial trainability challenges, as discussed in Section 1.2. This review asks which ansatz-design strategies most effectively preserve usable gradients in few-qubit QNNs and which interactions among those strategies remain unresolved. The term *few-qubit* refers here to 2–10 qubits, a deliberately restricted range chosen for controlled finite-shot evaluation and classical verification rather than as a

characterization of the full scale of current NISQ processors.

1.2 The barren plateau limitation

Despite the promise of QNNs, barren plateaus pose a major obstacle to trainability. Formally, a barren plateau is a region of the cost landscape in which the variance of the cost-function gradient vanishes exponentially with the number of qubits n: $Var\left[\partial_\theta C\right] \in O\left(b^{-n}\right), b>1$ (eq1).

Maintaining a fixed precision for typical gradient estimates can therefore require a number of circuit evaluations that grows exponentially with system size (3). McClean et al. showed this concentration for sufficiently random parameterized circuits. As such circuits approach highly scrambled ensembles, measured costs and gradients concentrate around their means; under a practical measurement budget, the optimizer may consequently receive no statistically reliable update direction.

Although barren plateaus were first identified in random deep circuits, subsequent work established several distinct concentration mechanisms. Global cost functions produced exponential gradient-variance decay under specified shallow-circuit assumptions (4); hardware noise induced concentration by driving states toward mixed distributions (5); and excessive entanglement has been connected to gradient suppression (6). Ragone et al. (7) provided a unified framework connecting barren-plateau behavior to a circuit's dynamical Lie algebra. These mechanisms can affect finite systems even when a 2–10-qubit sweep is too short to establish an asymptotic scaling class. Few-qubit evaluation must therefore measure finite-size gradient signal and estimator uncertainty directly rather than infer trainability from qubit count alone.

1.2.1 *Why gradient SNR is an operationally relevant metric in the few-qubit regime*

An important implication of the few-qubit regime is that the exponential scaling defined in Equation 1 generally cannot be established robustly from numerical results over such a restricted range. Adopting gradient SNR as a primary operational metric of gradient-based trainability is therefore not merely a substitution for an unavailable asymptotic measure. It follows from a chain of considerations that make estimator SNR directly relevant when only 2–10 qubits are available and every gradient estimate must be obtained under a finite measurement budget.

1.2.1.1 Step 1: Gradient-based optimization is constrained by gradient usability, not merely gradient existence

A foundational concern in variational optimization is not merely whether a gradient is mathematically nonzero, but whether its estimate is sufficiently informative to guide updates efficiently under the stated optimizer and measurement budget. Arrasmith et al. (8) showed that increasingly flat variational landscapes imposed severe precision and measurement requirements on both gradient-based and gradient-free optimization; their result did not establish a universal gradient-magnitude threshold. Surveys of practical VQA optimization have likewise documented failure modes involving gradient magnitude, estimator quality, and finite sampling without requiring

prior confirmation of an asymptotic exponential decay law (9,10). Tamiya and Yamasaki (11) used a gradient-norm test to adapt the shot count so that the estimated update direction maintained a controlled signal-to-noise relation. Collectively, these results support evaluating whether a gradient estimate remains usable for parameter updates while not implying that every stochastic optimizer requires a universal per-step or per-component SNR threshold.

1.2.1.2 Step 2: Finite-shot estimation introduces a practical uncertainty scale

Gradients in variational quantum circuits are not generally supplied to the optimizer as exact analytic values. For expectation-value costs and supported gate generators, they can be estimated through parameter-shift rules (12,13). Each finite-shot estimate is random: the optimizer observes an estimator of the derivative rather than the derivative itself. For a fixed estimator and shot-allocation rule, the standard deviation of sampling uncertainty scales as $O\left(N^{-1/2}\right)$ for measurement budget N (14). The proportionality constant is not universal; it depends on the prepared state, measured observable, shift coefficients, number and grouping of measurement terms, and whether N denotes shots per shifted circuit or a total budget distributed across all required circuits.

Finite sampling therefore creates a configuration-dependent uncertainty scale rather than an irreducible or universal numerical floor. The ratio of the estimator's repeated-shot mean to its sampling uncertainty is operationally relevant because it describes whether the signal supplied to the optimizer is statistically resolvable. For an unbiased estimator, this mean equals the exact derivative; for a biased end-to-end estimator, estimator SNR must be interpreted together with the exact-gradient reference, estimator bias, and sign agreement. Gradient SNR is not introduced here as a new statistic; the contribution is to place it at the center of few-qubit trainability evaluation while retaining exact-gradient landscape variance as a separate theoretical diagnostic (11). Recent work strengthens the practical connection between finite sampling, gradient measurement, and training cost. Kreplin and Roth (15) empirically showed that reducing finite-sampling variance in QNN outputs accelerated training and reduced the number of gradient-circuit evaluations required in their benchmarks, including a hardware demonstration. Chinzei et al. (16) proved a trade-off between gradient-measurement efficiency and expressivity for a broad class of deep QNNs and demonstrated in four-qubit numerical experiments that a structured ansatz could reduce training sample complexity while maintaining accuracy and trainability relative to parameter-shift baselines.

These results do not establish SNR as a universal convergence condition; they show that finite-sampling variance and gradient-measurement cost are substantive parts of the practical trainability problem rather than implementation details external to it.

1.2.1.3 Step 3: Few-qubit studies generally cannot establish asymptotic scaling from numerical sweeps alone

The exponential decay defined in Equation 1 is an asymptotic statement about behavior as $n \to \infty$. Robust inference of exponential scaling requires enough system sizes and sufficiently precise gradient-variance estimates to distinguish $O\left(b^{-n}\right)$ from $O\left(n^{-c}\right)$ for $c>0$ decay.

Distinguishing functional forms from a small number of noisy observations is a well-known challenge in empirical scaling and finite-size analysis (17,18). Within the 2–10-qubit window, finite-size corrections, initialization dependence, circuit-depth dependence, and measurement uncertainty can allow multiple candidate models to describe the same data with comparable support (9,10). For illustration, 2^{-n} and n^{-3} remain within one order of magnitude throughout $n \in \{2,\ldots,10\}$ and are nearly equal at $n=10$. This numerical coincidence does not prove that the two forms are statistically indistinguishable in every experiment, but it demonstrates why a short sweep can remain compatible with qualitatively different asymptotic interpretations. A defensible scaling claim would therefore require uncertainty estimates, explicit model comparison, and preferably theoretical support rather than a least-squares fit alone.

This distinction is reflected in the foundational barren-plateau literature. McClean et al. (3) established exponential gradient concentration analytically and used numerical results as finite-size confirmation. Cerezo et al. (4) similarly derived cost-function-dependent scaling under specified circuit-depth assumptions, while Holmes et al. (19) connected expressibility to gradient concentration through analytic bounds supplemented by finite-size numerics. These studies do not render few-qubit measurements uninformative; rather, they clarify what those measurements directly establish. In the absence of a supporting theorem, measurements across 2–10 qubits characterize finite-size gradient concentration, estimator uncertainty, and optimization behavior more directly than they determine the asymptotic scaling class (7). The same caution applies to any decay model fitted to a limited set of noisy system sizes: numerical evidence can support consistency with a theory but should not be treated as independent proof of the asymptotic form.

1.2.1.4 Synthesis

Because gradient-based optimization is affected by the quality of the estimates it receives (Step 1), those estimates are obtained under finite sampling with nonzero and configuration-dependent uncertainty (Step 2), and few-qubit numerical studies generally cannot establish asymptotic variance scaling by themselves (Step 3), pointwise estimator SNR provides an operational measure of one practically important aspect of trainability: whether the estimated update signal is distinguishable from its sampling uncertainty under a specified estimator and budget. This quantity can be evaluated more directly in a few-qubit experiment than an asymptotic decay class inferred from a short sweep. It does not replace exact-gradient variance as a theoretical barren-plateau diagnostic, establish estimator correctness, or impose a universal convergence threshold; instead, it addresses a distinct and immediate experimental question.

Two distinct sources of variance must therefore remain separate throughout the analysis. The first is the variance of the exact gradient across parameter space, $Var_{\theta}\left[g_k^{\text{exact}}(\theta)\right]$, which characterizes landscape concentration and is the quantity relevant to Equation 1. The second is the sampling variance of a finite-shot gradient estimator at a fixed parameter point, $Var_{\text{shots}}\left[\hat{g}_k(\theta) \mid \theta\right]$. The first concerns how the exact landscape changes as parameters or

initializations are varied; the second concerns how repeated measurements fluctuate while the parameter point is held fixed. Conflating these quantities would incorrectly imply that any strategy reducing landscape variance necessarily lowers the denominator of SNR, even though shot-estimator variance is determined by the measurement protocol and state-dependent outcome statistics.

For parameter θ_k, define the exact total derivative $g_k^{\text{exact}}(\theta) = \frac{\partial C(\theta)}{\partial \theta_k}$, and let $\hat{g}_k(\theta)$ denote the finite-shot gradient estimator supplied to the optimizer. Parameter-shift rules provided exact derivatives of supported quantum expectation values under ideal evaluation (12,13). For the residual hybrid architecture proposed in Section 5.2, the complete estimator is instead assembled by combining node-level parameter-shift estimates with the exact derivatives of the intervening classical maps through the chain rule. Its expectation over repeated shot realizations at the same parameter setting is $\mu_k(\theta) = E_{\text{shots}}\left[\hat{g}_k(\theta) \mid \theta\right]$.

The primary pointwise estimator SNR is then defined as $SNR_k^{\text{est}}(\theta) = \frac{|\mu_k(\theta)|}{\sqrt{Var_{\text{shots}}\left[\hat{g}_k(\theta) \mid \theta\right]}}$ (eq2).

When an exact statevector derivative or a prespecified high-precision reference is available, the complementary exact-gradient resolvability is $SNR_k^{\text{exact}}(\theta) = \frac{|g_k^{\text{exact}}(\theta)|}{\sqrt{Var_{\text{shots}}\left[\hat{g}_k(\theta) \mid \theta\right]}}$.

The estimator bias is $Bias_k(\theta) = \mu_k(\theta) - g_k^{\text{exact}}(\theta)$. For an unbiased estimator, $\mu_k = g_k^{\text{exact}}$ and the two SNR definitions coincide. For a biased estimator, SNR^{est} evaluates whether the mean update signal actually supplied to the optimizer is distinguishable from its finite-shot uncertainty, whereas SNR^{exact}, bias, and sign agreement evaluate whether that resolvable signal faithfully represents the exact derivative.

The expectation in μ_k is over repeated shot realizations at a fixed parameter point, not over randomly sampled parameter initializations. Positive and negative estimator means therefore do not cancel across the parameter space. This distinction is important because $\left|E_\theta[\mu_k]\right|$, $E_\theta[|\mu_k|]$, and $\sqrt{E_\theta[\mu_k^2]}$ are different quantities and answer different questions. Equation 2 evaluates the statistical resolvability of a particular total-gradient estimator component at the parameter point where the optimizer must act; it does not by itself establish that the estimator is unbiased or has the correct sign.

At the quantum-node level, the standard deviation of a fixed parameter-shift estimate scales as $O(N^{-1/2})$ under a fixed measurement protocol (14). For a fixed hybrid architecture and prespecified allocation of independent finite-variance shot batches, the assembled chain-rule estimator retains this leading-order dependence on the total measurement scale, but its proportionality constant also depends on downstream sensitivities, products of node Jacobians, observable grouping, and the allocation of shots across the computational graph. Increasing the measurement budget can therefore improve estimator SNR, but resolving

an exponentially small exact total-gradient signal still requires a correspondingly increasing number of measurements. Higher pointwise SNR^{est} indicates that the estimator mean is more reliably distinguishable from finite-shot uncertainty under the stated architecture, estimator, and budget; it does not establish that the mean is unbiased or directionally correct. Estimator SNR is therefore adopted as an operational diagnostic of gradient resolvability and sampling efficiency under a specified estimator and measurement budget. Persistently low SNR can slow, destabilize, or substantially increase the measurement cost of gradient-based optimization, whereas high SNR does not guarantee convergence or correctness because estimator bias, sign error, loss-landscape geometry, optimizer dynamics, parameter correlations, local minima, and hardware bias also play a role. Conversely, stochastic-gradient methods can converge with individually noisy estimates under suitable unbiasedness, variance, and step-size conditions (14); no universal per-component SNR threshold is asserted. Arrasmith et al. (8) support the broader claim that increasing landscape flatness creates severe optimization precision requirements, not a universal gradient magnitude or SNR at which a particular optimizer must succeed.

For empirical comparison, Equation 2 should be evaluated separately for each parameter and initialization using repeated, independent shot batches. In the factorial experiment, the primary estimator-SNR summaries are restricted to matched quantum-gate parameters present in every configuration. Gradients of the classical matrices, biases, and residual coefficients are reported separately and are not pooled into the primary estimator-SNR distribution because those parameters are not present in every condition and need not share a common scale. The matched quantum-gradient distribution should be reported rather than collapsed into the signed average $\left|E_\theta\left[\mu_k\right]\right|$. Recommended primary summaries include the median pointwise SNR^{est}, its interquartile range, the fraction of gradient components with $SNR^{est} \geq 1$, and the dependence of these summaries on circuit depth and shot budget. The fraction of gradient components with estimator SNR ≥ 1 is used only as a descriptive resolvability benchmark, indicating that the magnitude of the estimator mean is at least as large as one estimated standard deviation of its finite-shot sampling uncertainty; it is not treated as a necessary or sufficient threshold for optimization success. Where an exact or high-precision reference is available, SNR^{exact}, estimator bias, absolute bias, and the fraction of components satisfying $sign\left(\mu_k\right) = sign\left(g_k^{\text{exact}}\right)$ must also be reported. Exact-gradient landscape variance $Var_\theta\left[g_k^{\text{exact}}\right]$ is reported separately as a secondary measure. No universal numerical "noise floor" follows from N alone: even for a bounded observable, the variance depends on the state, observable decomposition, chain-rule sensitivities, shift coefficients, and allocation of shots among all measurements required by the computational graph.

Several assumptions underlie this framing. First, the 2–10-qubit range is adopted because it permits tractable near-term experiments and classical verification; it does not represent the full scale of existing NISQ processors, and conclusions may not generalize directly to larger systems. Second, pointwise estimator SNR is adopted as a trainability proxy because

distinguishability from finite-shot uncertainty is informative about update reliability and measurement cost under the stated estimator. It is neither a universal necessary nor sufficient condition for convergence, and the metric does not capture loss-landscape geometry, optimizer choice, parameter correlation, stochastic averaging across iterations, hardware bias, or task fidelity. Third, the three candidate strategies examined in Section 5—restricted entanglement, local cost functions, and residual connections—were selected because each has been evaluated in isolation (4,6,20), because they act at different structural levels, and because their joint behavior remains a nontrivial practical question. This selection does not imply that they form an optimal or complete set. Fourth, restricted entangling connectivity is assumed to preserve sufficient representational capacity for the local target task used in the proposed experiment. Final energy and fidelity must test that assumption, because a trainable but insufficiently expressive ansatz would not constitute a successful mitigation strategy.

1.3 Motivation for mitigating barren plateaus

Previous research has identified several mechanisms underlying barren plateaus and thereby guided mitigation efforts. Random entanglement generated during circuit evolution has been identified as one contributor (21). By controlling entanglement—for example, through initial register partitioning or restricted parameter ranges—prior work produced more favorable training landscapes in which gradients retained useful information. Dynamical-Lie-algebra frameworks have also connected circuit structure to barren-plateau behavior and supplied predictive tools for assessing trainability before a design was implemented (22,23).

Given the limitations posed by barren plateaus, mitigation is important for the practical deployment of QNNs. Prior work has identified several contributors: random or excessive entanglement generated during circuit evolution (3,6), global cost functions that can induce exponential gradient decay even in shallow circuits (4), and hardware noise that can drive states toward maximally mixed distributions (5). Depending on circuit structure, task, and noise level, these mechanisms have substantially degraded optimization even before large-system asymptotics were numerically accessible (9). Four broad mitigation categories have emerged. Cerezo et al. (4) showed that local cost functions yielded at worst polynomial gradient-variance decay under specified logarithmic-depth assumptions, whereas global observables produced exponential decay in the analyzed setting. Kashif and Al-Kuwari (20) proposed Residual Quantum Neural Networks (ResQNets), which incorporated shortcut connections intended to preserve information and gradient flow. He et al. (24) earlier showed that classical residual connections improved deep-network optimization; ResQNets adapted this architectural principle to hybrid quantum models. Initialization methods (25–27) have addressed flatness near the start of training, while zero-noise extrapolation and probabilistic error cancellation have targeted hardware error. Structured ansatz design has addressed gradient concentration through circuit connectivity, depth, symmetry, and expressibility.

Each strategy may affect the exact gradient signal and finite-shot estimator uncertainty

through different pathways. Restricted entanglement may preserve gradient magnitude by reducing state scrambling, but it may also reduce representational capacity and weaken task-relevant gradients. Local cost functions can strengthen local gradient components while discarding information about global correlations. Residual connections may preserve signal across depth, but their implementation can alter the measured features, effective circuit structure, or accumulation of hardware error. None of these structural effects should automatically be described as lowering the SNR denominator: the denominator in Equation 2 is the variance of the finite-shot estimator at a fixed parameter point and depends on the observable, state, and measurement allocation. The combined effect of multiple strategies on estimator SNR is therefore nontrivial and cannot be assumed to be additive. The following sections examine each strategy and its trade-offs, with particular emphasis on ansatz design as the most structurally fundamental of the four approaches.

1.4 Overview of mitigation strategy categories

Barren-plateau mitigation strategies can be grouped into four broad approaches: ansatz design, parameter initialization, cost-function selection, and error management. This section summarizes their contributions, limitations, and relationships.

1.4.1 *Ansatz design*

Ansatz design plays a central role in QNN trainability because it determines circuit depth, connectivity, symmetry, expressibility, and the states on which the cost is measured. Prior work showed that sufficiently expressive circuits approaching unitary 2-design behavior could exhibit strongly concentrated gradients (19). In contrast, ansätze inspired by tree tensor networks or quantum convolutional neural networks resisted barren plateaus in the settings analyzed, with gradient magnitudes or variances exhibiting polynomial rather than exponential scaling (28–30). Other mitigation categories can interact with or compensate for ansatz choices, but they do not remove this structural role. Under the SNR framing in Section 1.2.1, structured ansatz design acts most directly on exact-gradient magnitude by limiting scrambling; for an approximately unbiased estimator, this change appears in the estimator-mean numerator. Ansatz design may also change finite-shot estimator variance indirectly by altering the prepared state or measurement requirements, but that effect must be evaluated rather than inferred from expressibility alone.

1.4.2 *Initialization strategies*

Proper parameter initialization is important for avoiding barren plateaus near the start of training, but it does not prevent flat regions from emerging later. Identity-block initialization sets each circuit block to evaluate initially as the identity, keeping the starting state near a controlled reference and facilitating early gradient flow (25). Transfer-learning-inspired initialization reuses parameters from smaller or related tasks to improve convergence (26). Beta-distribution initialization samples parameters from a fitted distribution to reduce concentration near training onset (27). Initialization is therefore complementary rather than substitutive: a well-initialized circuit can still enter a barren plateau if its subsequent dynamics or structure are insufficiently constrained.

1.4.3 *Cost function selection*

The cost function strongly influences gradient behavior and interacts with ansatz structure. Cerezo et al. (4) showed that global observables produced exponentially vanishing gradient variance in the analyzed setting, whereas local cost functions yielded at worst polynomial decay when ansatz depth grew no faster than logarithmically with qubit count. In SNR terms, this result concerns signal-side scaling; it does not establish that a pointwise gradient exceeds the finite-shot uncertainty of an unspecified estimator under an unspecified shot budget. Resolvability must be evaluated for the prepared state, observable decomposition, and measurement allocation used in the experiment.

Cost-function selection is therefore distinct from ansatz design, but local objectives may omit global information and create a trade-off between gradient usability and task fidelity.

1.4.4 *Error management*

Hardware noise can induce barren-plateau behavior that circuit design alone cannot fully address, making error management a necessary complement to structural strategies. Zero-noise extrapolation evaluates circuits at deliberately varied noise levels and extrapolates toward a zero-noise expectation (31,32). Probabilistic error cancellation characterizes a device noise channel and samples a quasiprobability representation intended to cancel its effect statistically (32). These methods seek to reduce hardware-induced distortion, although they do not automatically reduce finite-shot estimator variance (33). Their sampling overhead can grow rapidly or exponentially with circuit size, limiting scalability and favoring combinations with structural mitigations that reduce depth and noise exposure.

2. Ansatz design for mitigating barren plateaus in few-qubit QNNs

A QNN ansatz typically alternates parameterized single-qubit rotations with two-qubit entangling gates. A common entangling operation is the CNOT gate, in which one qubit controls whether a second qubit is flipped. Repeated in layers, such gates spread correlations across the register and expand the set of states the circuit can represent (19). Expressibility generally increases with depth and entanglement density, but this benefit can reduce trainability: circuits approaching unitary 2-design behavior generate broadly distributed states for which gradient information becomes strongly concentrated (19).

Structured ansatz design is hypothesized to improve QNN training by preserving meaningful gradient information rather than allowing the landscape to flatten. Few-qubit systems permit direct classical verification and controlled finite-size characterization, but their limited size does not establish the absence of barren-plateau mechanisms (3,19). Material gradient suppression can still arise from noise, cost-function choice, or entanglement even when asymptotic scaling is not identifiable. Distinct strategies have therefore emerged to control ansatz structure through different mechanisms.

2.1 Hardware-efficient ansätze

Hardware-efficient ansätze (HEAs) were introduced to limit compilation overhead through shallow layers and native device connectivity (34). A typical HEA alternates

parameterized single-qubit rotations with entangling gates that follow the hardware coupling graph. This construction reduces implementation overhead, but the designation "hardware-efficient" does not itself guarantee favorable gradient scaling.

Shallow native-gate circuits can reduce noise exposure and delay correlation spreading, although the effect depends on the task, initialization, connectivity, and depth (19). Limiting depth can reduce accumulated hardware error (5), while restricting connectivity relative to a highly random circuit can slow state scrambling (3). Neither benefit follows from the "hardware-efficient" label alone. Under the SNR framing, a shallower circuit may preserve exact-gradient magnitude and reduce hardware-induced distortion; when estimator bias is negligible, the signal-side benefit appears in the estimator-mean numerator. The finite-shot denominator remains determined by state-dependent measurement statistics and shot allocation and must therefore be measured rather than assumed to decrease with circuit depth. Park and Killoran (35) derived trainability conditions for a Hamiltonian variational ansatz approximable by local-Hamiltonian time evolution. That result applies to the structured HVA under its stated parameter conditions and should not be generalized to arbitrary HEAs. Moreover, reducing depth can make a generic HEA insufficiently expressive. Adding layers to recover expressibility can then increase entanglement and noise exposure, partially reversing the trainability benefit.

2.2 Problem-inspired ansätze

A problem-inspired ansatz tailors circuit structure to a specific task rather than relying on a generic construction (9). Canonical examples include QAOA-inspired circuits for combinatorial optimization, which alternate cost and mixer unitaries that encode the objective, and quantum convolutional neural networks (QCNNs), which use structured convolutional and pooling layers that preserve spatial locality. Subsequent work has examined the conditions under which the absence of barren plateaus in QCNNs holds and whether those guarantees extend beyond the specific architectures analyzed in Refs. (29,30). The guarantee should therefore not be generalized without task- and architecture-specific validation.

Although QAOA is a variational optimization algorithm rather than a QNN, it shares the relevant feature of a problem-structured parameterized circuit. Larocca et al. (10) treated QAOA-style circuits within the broader variational-ansatz framework, and the same convention is adopted here. One structural distinction remains important: QAOA uses two parameters per layer regardless of qubit count, whereas general QNN ansätze often increase parameter count with system size. QAOA's parameter efficiency therefore need not transfer directly to QNN settings. Tree-tensor-network ansätze likewise enforce hierarchical entanglement structure and exhibit favorable trainability scaling in the settings analyzed (28).

By aligning circuit structure with known symmetries, locality, or problem operators, problem-inspired ansätze can restrict exploration to a task-relevant subspace. Specific QCNN and tensor-network constructions exhibited favorable gradient scaling under their stated assumptions (28,29). Under the SNR framing, such structure may preserve task-

relevant exact-gradient magnitude, but neither larger pointwise estimator SNR nor lower measurement cost follows without specifying the observable and estimator. The benefit is therefore architecture- and task-specific rather than a general property of problem-inspired circuits.

This specificity is also a limitation. A problem-inspired ansatz is tuned to particular structure, and adaptation to a different task may require substantial redesign. Such circuits are also less useful when the relevant structure is unknown or difficult to encode in quantum operations, reducing their generality relative to hardware-efficient designs.

2.3 Comparative analysis and few-qubit implications

Hardware-efficient and problem-inspired constructions impose different and potentially interacting constraints. HEAs prioritize native implementation and shallow depth, whereas problem-inspired ansätze encode task structure through symmetries, locality, or problem operators. Either approach can be trainable or untrainable depending on initialization, depth, cost function, and noise; neither is universally superior.

Few-qubit systems permit controlled finite-size comparisons of these approaches but do not determine their asymptotic trainability (3,19). At this scale, parameter count, gate count, entangling connectivity, and classical verification can be controlled explicitly. The relevant comparison is therefore empirical: whether each architecture preserves resolvable gradients and reaches the common task target under a matched resource budget.

2.3.1 *Design recommendation*

When known problem structure maps onto an architecture for which barren-plateau absence has been formally demonstrated—specifically, translationally invariant QCNN circuits (29,30) or hierarchical tree-tensor-network ansätze (28)—those architectures demonstrated more favorable gradient scaling than generic random initialization in the settings analyzed. This advantage has not been established for problem-inspired ansätze as a general class. In particular, the QCNN guarantee depends on the stated circuit structure and need not hold for problem-inspired designs lacking the same symmetry (29). Other problem-inspired constructions therefore require empirical validation in the specific few-qubit task. When relevant problem structure cannot be encoded in a circuit family with demonstrated trainability properties, a shallow hardware-efficient ansatz provides a practical controlled baseline rather than a guarantee of barren-plateau mitigation. Interactions with local costs, residual connections, and initialization methods remain open questions, as discussed in Section 4.

3. Limitations and technical challenges

Despite substantial progress in barren-plateau mitigation, several barriers continue to limit practical QNN deployment. These barriers include circuit-design trade-offs, hardware constraints, and fundamental optimization difficulties.

3.1 Trade-offs in ansatz and cost function design

Ansatz and cost-function design remain central challenges for QNNs. Highly expressive ansätze may provide the representational capacity required for complex tasks, but they can also increase gradient concentration through

excessive entanglement and state scrambling (19). Restrictive or hardware-efficient ansätze can reduce trainability barriers yet fail to represent the target function. Similarly, global cost functions are natural for many quantum applications but can induce vanishing gradients even in shallow circuits, motivating local or problem-informed alternatives (4). These tensions create a persistent expressibility–trainability trade-off. Specific structured ansätze exhibited polynomial rather than exponential gradient-variance scaling under their stated assumptions (4,28,29), $Var\left[\partial_\theta C\right] \in O\left(n^{-c}\right), c>0$ (eq3).

Polynomial gradient-variance decay is asymptotically more favorable than exponential decay, but neither form alone determines whether a gradient is resolvable in a fixed 2–10-qubit experiment. Constants, circuit depth, observable choice, initialization, and measurement budget remain consequential. When the exact-gradient scale decays exponentially with qubit count, maintaining fixed estimator SNR generally requires a measurement budget that also grows exponentially (3,4). A small fixed system may still be measurable with sufficient sampling; the obstacle is unfavorable resource scaling rather than absolute impossibility at every finite size.

3.2 Noise and hardware limitations

Near-term quantum devices are noisy, and hardware errors can exacerbate barren-plateau behavior by driving states toward mixed distributions and flattening the optimization landscape (5). The degree of gradient suppression depends on the noise channel, circuit depth, and measured observable. Hardware-connectivity constraints can also force additional routing or altered entangling structures, introducing further compromises in ansatz design (34).

3.3 Optimization landscape and training challenges

Variational cost landscapes can be non-convex and sensitive to initialization, producing unstable convergence or local optima (36). Even a few-qubit system can supply an unreliable update when its exact gradient is small relative to finite-shot or hardware-induced uncertainty. Arrasmith et al. (8) quantified the precision burden created by flat landscapes, while parameter-shift and stochastic-optimization studies formalized the repeated finite-shot evaluations required to estimate expectation values and gradients (12–14).

Expressibility must also be distinguished from generalization. A restrictive ansatz can fail to represent the target, whereas an overly flexible quantum model can overfit a limited training set (37). Problem-inspired structure may improve inductive bias for a matched task but reduce transfer to tasks requiring different correlations. The resulting design problem remains unresolved: circuits that are too shallow may lack representational power, while deeper circuits can introduce additional concentration and hardware noise.

Table 1. Comparison of barren-plateau mitigation strategies. The SNR interpretations apply to each intervention in isolation; combined effects remain empirical and must be evaluated under matched measurement budgets.

Strategy	Literature-supported mechanism	Operational SNR interpretation	Principal limitation or control
Ansatz design	Limits scrambling through depth, connectivity, symmetry, or problem structure.	Primarily changes exact-gradient magnitude; finite-shot variance remains state- and measurement-dependent.	Expressibility–trainability trade-off; no universal architecture rule (19).
Parameter initialization	Avoids highly concentrated regions near training onset.	Can improve early gradient usability, but the benefit may change during optimization.	Does not prevent later flatness; initialization distribution must be matched (25,26).
Cost-function selection	Local observables yielded more favorable gradient-variance scaling under specified depth assumptions.	Can preserve signal-side scaling; resolvability still depends on estimator variance and shot allocation.	Local objectives may omit global information; final energy and fidelity must be checked (4).
Error management	Reduces hardware-induced distortion or extrapolates toward lower-noise expectations.	May improve practical estimator quality but can add substantial sampling variance and overhead.	Sampling cost can grow rapidly; ideal, finite-shot, and noisy conditions must be separated (32,33).

4. Gaps in literature

Although meaningful progress has been made in ansatz design, important gaps remain in the study of combined mitigation strategies, the limits of the expressibility–trainability trade-off, and the theoretical foundations for task-specific ansatz selection. The surveyed studies generally examined depth, entanglement structure, cost-function choice, and residual design as separate interventions; they did not cross the three components considered here in a complete factorial comparison (10).

4.1 Unexamined combinations of mitigation strategies

Although many studies have examined expressibility control, entanglement restrictions, local costs, or residual connections, their combined effects at the ansatz level remain largely unresolved. This gap matters because entanglement, cost-function support, hardware noise, and circuit depth can influence the same NISQ experiment simultaneously. An intervention targeting one pathway may be offset by another, yet the cited studies have not systematically estimated the corresponding interaction terms for the three components considered here. For example, a local cost can preserve favorable signal-side scaling while hardware noise drives the state toward a mixed distribution, hence practical finite-shot benefit cannot be inferred from the ideal local-cost result alone. Residual QNNs split a conventional architecture into quantum nodes connected through residual pathways (20), while classical splitting partitioned an N-qubit ansatz into smaller $O(\log N)$-qubit circuits (38). These approaches were developed separately, and neither isolated result determines the entanglement×cost×residual interactions proposed here.

Residual quantum architectures demonstrated improved training behavior relative to plain

QNNs in the configurations tested (20), but those comparisons did not cross the residual intervention with both local-cost and entanglement choices. It therefore remains unknown whether combining a residual architecture with a local cost—which yielded more favorable gradient scaling under shallow or logarithmic-depth assumptions (4)—would produce an additional gain, a diminishing return, or a reversal under matched finite-shot resources. Classical splitting was likewise evaluated as a distinct architecture rather than as one factor in a unified comparison (38).

These gaps have practical significance. In few-qubit NISQ experiments, identifying mutually reinforcing combinations could support more efficient designs without assuming that gradient usability and expressibility improve together. Future work should therefore use controlled factorial studies to isolate and combine mitigation strategies and to determine which pairings are effective under specified depth, task, measurement, and noise conditions.

Studying combinations may appear to be ordinary ablation practice. The non-obvious contribution proposed here is not the recommendation to combine interventions, but the use of an operational SNR framework to separate exact-gradient signal from finite-shot uncertainty and to formulate interaction questions before data are analyzed. Existing few-qubit mitigation studies most often compared one intervention with an unmitigated baseline. Larocca et al. (10) surveyed strategies whose evidence bases were largely developed separately; Cerezo et al. (4) established local cost-function results under specified ansatz-depth conditions without simultaneously testing residual architecture and entanglement restriction; and Kashif and Al-Kuwari (20) evaluated residual quantum architectures without a complete factorial comparison across cost-function and entanglement choices. This pattern does not show that prior literature assumed additivity or claimed that isolated gains could be multiplied. It does leave the signs and magnitudes of the corresponding interaction terms uncharacterized.

The SNR framing organizes this gap by separating two questions. First, how does each intervention alter the exact gradient $g_k^{\text{exact}}(\theta)$? Second, how does it alter the estimator mean $\mu_k(\theta)$, finite-shot variance, and bias relative to that exact signal? Gradient variance across parameter space remains an important barren-plateau diagnostic, but it does not identify whether a change in practical gradient usability arose from a larger exact derivative, a different measurement variance, or both. Separating these quantities permits strategy combinations to be compared under a fixed total measurement budget rather than through an undifferentiated notion of "larger gradients." Non-additivity thereby becomes an explicit statistical estimand rather than a qualitative afterthought.

The strongest mechanistic motivation concerns restricted entanglement paired with a local cost. Restricted entanglement changes how parameter variations propagate into the reduced state of a measured subsystem, while the local cost determines which components of that reduced-state change contribute to the gradient. The interventions therefore act on overlapping determinants of the exact local gradient, even though the local observable does not restrict the

circuit's state distribution. If the entangling architecture has already preserved derivative structure relevant to the measured subsystem, replacing a global cost with a local one may yield a smaller marginal benefit than when the local cost is introduced in isolation. This reasoning motivates testing the pair's exact-gradient interaction, with sub-additivity retained only as a prespecified interpretive expectation. It does not establish the outcome, determine the sign of the shot-variance change, or imply that the pair is harmful. The proposed factorial experiment is required to determine whether the overlap produces a measurable interaction.

Equivalent directional confidence is not assigned to interactions involving residual connections. Residual pathways may preserve information across depth, alter effective feature correlations, or change downstream sensitivities, but the available literature does not provide a comparable structural reason to predict a universal interaction direction with either entanglement restriction or local-cost choice. Those pathways are therefore treated as open pairwise or depth-dependent questions. The framework is intentionally asymmetric in interpretation: one pair has a mechanistically motivated expected direction, but all confirmatory interaction tests remain two-sided. This distinction prevents mechanisms with different evidentiary support from being presented as equivalent.

4.2 Theoretical open questions

Even within ansatz design, several theoretical questions remain unresolved, limiting principled design choices and encouraging trial-and-error selection.

One key question concerns the trade-off between ansatz expressibility and trainability: how expressive could an ansatz become before gradient concentration made optimization impractical? Holmes et al. (19) established a formal connection between expressibility—how closely an ansatz approaches a unitary 2-design, as defined in Section 1.4.1—and gradient concentration, showing that more expressive ansätze tended to produce flatter landscapes. A 2-design reproduces the first two statistical moments of the Haar measure, so relevant gradient information becomes strongly concentrated across broadly distributed states (3,19). This relationship has primarily been characterized through asymptotic theory and finite-size confirmation. In few-qubit systems, the threshold at which additional expressibility begins to impair trainability remains poorly resolved. Designers are therefore left without quantitative guidance on how much expressibility a particular task can support.

A second open question concerns the effects of circuit symmetries and structural constraints on few-qubit gradient behavior. Dynamical-Lie-algebra analyses have shown that controllability, reachable subspaces, and circuit symmetries can govern gradient concentration (7,22,23). Those results supply structural diagnostics but not a finite-shot map of how a particular restriction changes pointwise estimator SNR across depths, observables, and shot budgets in a 2–10-qubit implementation. Developing such maps would connect asymptotic or algebraic trainability theory to experimentally measurable design choices.

A third gap involves the trade-off between circuit depth and parameter density per layer. The literature has not established a general rule for choosing between a shallow circuit with more trainable rotations per layer and a deeper circuit with fewer parameters per layer under matched gate and measurement budgets. Depth can increase noise accumulation and correlation spreading, whereas parameter density can alter expressibility, initialization sensitivity, and the number of gradients requiring estimation. Residual pathways add an implementation-dependent factor because they can change information flow and classical parameter count without necessarily changing quantum depth (20). Resolving this trade-off would provide more actionable guidance than treating "shallow" or "parameter-efficient" as sufficient trainability criteria.

Together, these gaps create a common practical problem: circuit designers must combine approaches whose joint estimator-SNR behavior cannot be inferred from isolated characterizations. The expressibility–trainability threshold is poorly resolved in small Hilbert spaces; the finite-shot consequences of symmetry restrictions have not been mapped across depth and measurement budget; and the depth–parameter-density trade-off under residual connections remains unresolved. These limitations become especially consequential when multiple strategies are deployed simultaneously. Section 5 therefore proposes a falsifiable framework that treats interaction as an open empirical question, uses estimator SNR under finite sampling as the primary operational outcome, and supplies a complete factorial design for evaluation.

5. A testable framework for evaluating mitigation strategy interactions in few-qubit QNNs

The gaps identified in Section 4 share a common feature: existing mitigation strategies have generally been characterized in isolation, leaving their combined effects insufficiently resolved. The proposed framework does not assume that combining individually useful components will improve trainability. Instead, it creates a controlled design in which each main effect, each pairwise interaction, and the three-way combination can be evaluated under a common task, matched quantum-gate parameterization, matched initialization seeds, and equal total finite-shot measurement budgets. Residual-specific classical parameters are treated as part of the residual intervention and are reported separately rather than described as matched across all configurations. Estimator SNR supplies the primary operational outcome, while exact-gradient magnitude and variance, estimator bias and sign agreement, convergence, task fidelity, and resource cost remain necessary complementary measures.

Three candidate components are considered. A restricted entangling architecture is intended to slow indiscriminate correlation spreading and reduce state scrambling (6). A local cost function is intended to preserve locally informative gradient components under circuit-depth conditions in which a global objective may concentrate (4). A residual architecture is intended to preserve information and gradient flow across successive circuit blocks (20). Each component was motivated independently, but their combined behavior cannot be inferred from isolated comparisons because they can alter

overlapping stages of state preparation, cost evaluation, and gradient estimation.

5.1 Hypothesized interaction pathways between mitigation strategies

The three pathways below are assigned intentionally different evidentiary levels. Pathway A, restricted entanglement × local cost, has a structural basis in the exact local-gradient identity developed in Section 5.1.1; it motivates a prespecified test of the exact-gradient interaction and a directional interpretive expectation, not a one-sided derived prediction. Because the identity does not determine finite-shot variance, the complete estimator-SNR interaction receives no directional prior. Pathways B and C involve residual connections and rest on mechanistic reasoning without an equivalent derivation; they are therefore framed as non-directional or depth-dependent questions. All three pathways remain conjectures that motivate four falsifiable hypotheses (H1–H4, formally defined in Section 5.3), rather than conclusions established by this review.

5.1.1 *(A) Restricted entanglement × local cost functions*

Restricted entangling connectivity and a local cost both aim to preserve useful gradient information, but they intervene at different stages. The entangling architecture changes the state $\rho(\theta)$ and therefore how a parameter perturbation propagates through the circuit. The cost function does not restrict which states the circuit can prepare; it changes which features of the prepared state are scored and which components of the state derivative contribute to the gradient. The interventions should therefore not be described as restricting the same state distribution, although their effects can overlap at the measured derivative.

Let a local objective be written as a weighted sum of fixed observables $C_{loc}(\theta)=\sum_m w_m Tr\left[\rho_{S_m}(\theta) L_{S_m}\right]$, where L_{S_m} is supported on subsystem S_m, and $\rho_{S_m}=Tr_{\overline{S_m}}[\rho]$. Its exact gradient is,

$$g_k^{\text{exact}}(\theta)=\sum_m w_m Tr\left[\frac{\partial \rho_{S_m}(\theta)}{\partial \theta_k} L_{S_m}\right]. \quad \text{(eq4)}$$

Equation 4 is an exact identity for the sum-of-local-observables objective used in the proposed TFIM comparison. It is neither a decomposition of shot-noise variance nor, by itself, an interaction coefficient. It nevertheless identifies a pathway through which overlap may occur. A restricted entangling architecture can alter each reduced-state derivative $\partial \rho_{S_m}/\partial \theta_k$ by changing how parameter information spreads into the complement of S_m, while the local objective selects and weights the derivative components contracted with the observables L_{S_m}. The terms in Equation 4 can reinforce or cancel one another, so the identity does not imply a universal interaction sign. If the restricted architecture has already preserved locally relevant derivative structure, the local objective may provide a smaller marginal signal gain than it would under the baseline architecture; conversely, a restriction that blocks target correlations can reduce both exact-gradient magnitude and final task performance.

This overlap motivates a prespecified, two-sided test of the exact-gradient interaction. Sub-additivity is retained as an interpretive expectation, not as a consequence proved by

Equation 4. The finite-shot denominator must be measured independently because changes in the observable, state, or measurement allocation can either increase or decrease estimator variance. Equation 4 therefore supplies no directional prediction for the complete estimator-SNR interaction. Whether a signal-level interaction exists, and whether it persists, disappears, or reverses after sampling uncertainty is included, remains an empirical question.

5.1.2 *(B) Residual connections × entanglement restrictions*

Residual connections are intended to preserve information across depth by introducing shortcut paths between circuit blocks (20). Depending on their implementation, these paths may reduce the effective sequence of transformations traversed before information reaches the cost function. They could reinforce restricted entanglement by preserving useful signal without requiring additional deep quantum evolution, or they could partially undermine the restriction by reintroducing correlated features through a classical pathway or changing the dimensionality of the representation supplied to later blocks.

Unlike Pathway A, this interaction does not follow from Equation 4, and no structural argument currently favors sub-additivity over super-additivity. The two interventions may therefore interact non-independently, with direction determined by the residual implementation, circuit depth, and task. A complete factorial comparison is necessary because the fully combined configuration alone cannot identify whether a change arose from this pair or from the local cost.

5.1.3 *(C) Residual connections × local cost functions*

Residual architectures are expected to be most useful in deeper circuits, where information and gradients must propagate through more blocks. Local-cost guarantees, however, are generally derived for shallow or logarithmic-depth regimes (4). Their benefits may therefore peak at different depths: locality may provide its strongest protection before residual connections become necessary, while residual paths may become more useful after depth has weakened the advantage of the local objective. The interaction may consequently change in magnitude or sign as depth increases.

This pathway cannot be identified independently unless the ablation includes a local-cost×residual configuration without the entanglement restriction. Omitting that condition would force H4 to be inferred from the fully combined model and would confound the pairwise interaction with the third intervention. The complete design therefore includes the missing pair and treats circuit depth as an experimental factor.

Taken together, these pathways show why the three strategies should be analyzed as components of a coupled system rather than added informally as independent improvements. They also define the limited scope of the proposed theory: one pair receives a directional interpretive expectation at the exact-signal level, whereas the confirmatory test is two-sided and all complete estimator-SNR interactions remain open. The experiment is designed to reject, retain, or refine these priors.

5.2 Proposed experimental design

To evaluate all main and interaction effects, the proposed 2^3 factorial ablation comprises eight configurations:

[1] Baseline: a hardware-efficient ansatz with no additional mitigation.

[2] Restricted entanglement only: a fixed nearest-neighbor brick-layer entangling pattern, with the quantum-gate parameterization, entangling-gate type and count, trainable rotation count, and nominal layer count matched to the baseline.

[3] Local cost only: a local objective anchored to the same target state as the global objective, using the baseline ansatz.

[4] Residual connections only: shortcut connections added between ansatz blocks, using the global objective and baseline entangling architecture.

[5] Restricted entanglement + local cost: configurations [2] and [3] combined. This configuration is the primary test of H1 and H2.

[6] Restricted entanglement + residual connections: configurations [2] and [4] combined. This configuration is the primary test of H3.

[7] Local cost + residual connections: configurations [3] and [4] combined. This configuration is the primary test of H4.

[8] Combined: all three components active simultaneously, enabling evaluation of the three-way interaction and the exploratory combined-performance questions.

The eighth configuration is necessary rather than optional. Three binary interventions require one baseline condition, three single-component conditions, three pairwise conditions, and one three-component condition for a complete factorial design. Omitting the local-cost× residual pair would prevent independent estimation of that interaction. Including all eight conditions also permits formal analysis through factorial regression and normalized interaction indices rather than informal comparison of raw estimator-SNR values.

5.2.1 *Reference task and cost functions*

The reference task is preparation of the ground-state vector ψ_0 of a four-site, open-boundary transverse-field Ising model with Hamiltonian, $H = -J\sum_{i=0}^{2} Z_i Z_{i+1} - h\sum_{i=0}^{3} X_i$, (eq5), where $J=1$ and $h=0.5$. The task contains short-range correlations, is tractable on four qubits, and permits the exact target state and ground-state energy to be obtained by classical diagonalization. The system is not intended to establish asymptotic scaling; it provides a controlled setting in which the finite-shot interaction framework can be implemented and validated.

A valid comparison between global and local costs must hold the physical target fixed while preserving a well-defined finite-shot gradient estimator. Both proposed objectives are therefore anchored to the same target ground state. Let ρ_0 denote the projector onto ψ_0 and let $\rho(\theta)$ denote the prepared state. The global target-state infidelity is, $C_{\text{global}}(\theta) = 1 - Tr\left[\rho_0 \rho(\theta)\right]$, (eq6a), while the local objective is the normalized expectation value of the TFIM Hamiltonian, $C_{\text{local}}(\theta) = \frac{Tr\left[H\rho(\theta)\right] - E_0}{E_{max} - E_0}$. (eq6b), where E_0 and E_{max} are the minimum and maximum eigenvalues of H, obtained by classical diagonalization; for $J=1$ and $h=0.5$,

$E_0 \approx -3.4270$ and $E_{max} \approx 3.4270$. Equation 6b is a sum of fixed one- and two-qubit observables and is therefore local in the measurement-support sense used here. It has the same nondegenerate ground-state target as Equation 6a, reaches zero at that target, and admits standard parameter-shift estimation for the Pauli-generated rotations specified below. The two objectives nevertheless define different landscapes: Equation 6a measures global target-state infidelity, whereas Equation 6b aggregates locally measurable energy terms. Every configuration must therefore report both final TFIM energy and global fidelity. The global projector can be evaluated directly in statevector simulation or through a prespecified overlap circuit on hardware; its additional measurement or circuit cost must be included in matched resource accounting. This construction avoids treating a nonlinear tomography-derived mixed-state fidelity as though its final scalar value automatically obeyed a standard two-shift expectation-value rule. Accordingly, the estimated cost-function effect applies to this specific contrast between normalized TFIM energy and global target-state infidelity. It must not be attributed to measurement support alone, because the objectives also differ in spectral weighting, term aggregation, and measurement decomposition.

5.2.2 *Circuit controls and entanglement diagnostics*

The baseline ansatz consists of alternating R_y rotation and entangling layers, with the confirmatory depth sweep fixed at $L \in \{1,2,3,4,6\}$. Every layer applies one independently parameterized R_y rotation to each of the four qubits, followed by exactly four logical CNOT gates. In the baseline condition, every entangling layer uses the sequential directed-ring schedule $CX_{0\to1}, CX_{1\to2}, CX_{2\to3}, CX_{3\to0}$, in that order. This schedule contains a causal path spanning the complete register within one entangling layer. In the restricted condition, odd-numbered layers use $CX_{0\to1}, CX_{2\to3}, CX_{1\to0}, CX_{3\to2}$, and even-numbered layers use $CX_{1\to2}, CX_{3\to0}, CX_{2\to1}, CX_{0\to3}$.

The restricted schedule acts only within two disjoint pairs during any single layer; the pair partition changes between odd and even layers, hence circuit-wide propagation requires multiple layers rather than occurring within one layer. For the primary comparison, barriers are inserted between the four listed CNOTs in both conditions so that logical CNOT count and two-qubit depth are each fixed at four per entangling layer. CNOT direction, order, trainable rotation count, quantum-gate parameterization, initialization ranges, and nominal layer count are fixed as written. A barrier-free transpilation that exploits the restricted schedule's parallelism may be reported only as a secondary efficiency analysis. On hardware, native-gate decomposition, routing overhead, duration, and error rates are reported separately; if the prescribed logical schedules cannot be implemented with comparable compiled resources, the hardware comparison is labeled as a compound schedule-and-compilation intervention rather than attributed solely to entanglement restriction.

Area-law and volume-law behavior are scaling concepts that cannot be demonstrated in a single four-qubit system. The intervention is therefore described as a restricted nearest-neighbor, fixed-

depth entangling architecture rather than as evidence of area-law scaling. Bipartite entanglement entropy, reduced-state purity, or both are measured across the available cuts to determine whether the restricted architecture produces less subsystem mixing than the baseline without extrapolating an unobserved scaling law.

A reproducible residual implementation is prespecified. Each quantum block ℓ is treated as a differentiable quantum node, $z^{(\ell)} = q_\ell\left(x^{(\ell)}, \theta^{(\ell)}\right), z_j^{(\ell)} = Tr\left[Z_j \rho_\ell\left(x^{(\ell)}, \theta^{(\ell)}\right)\right]$, where $x^{(\ell)}$ contains the encoded inputs to the block and $\theta^{(\ell)}$ contains its trainable quantum-gate parameters. In the baseline architecture, the encoding angles for the next block are $x^{(\ell+1)} = W_\ell z^{(\ell)} + b_\ell$. In the primary residual condition they are $x^{(\ell+1)} = W_\ell z^{(\ell)} + b_\ell + \gamma z^{(\ell-1)}$, where the shortcut gain is fixed at the prespecified nonzero value $\gamma = 1$ and feature dimensions are held fixed. The baseline is equivalently represented by $\gamma = 0$. Because the shortcut is active at the matched initial parameter point, the primary analyses can identify residual main and interaction effects before optimization begins. Kashif and Al-Kuwari introduced the broader hybrid residual principle on which this shortcut is based (20). The selected implementation adds no trainable parameters, no quantum gates, and no feature observables beyond those measured by the matched baseline; W_ℓ and b_ℓ are parameterized and initialized identically in residual and non-residual conditions. A gain sensitivity analysis with prespecified nonzero values, such as $\gamma \in \{0.5, 1\}$, may be reported as exploratory. A separate trainable-gain variant initialized at zero may be studied longitudinally, but it is not used for the matched-initialization tests H3 or H4 because a zero gain would make the shortcut inactive at the primary analysis point. Alternative residual constructions must be treated as separate implementations because adding gates, changing measured features, adding trainable parameters, or requiring additional circuits would alter both resource accounting and the interaction mechanism.

Because an upstream quantum parameter affects the final cost through subsequent re-encoding operations and quantum blocks, its total derivative is evaluated through the chain rule rather than by applying a two-shift formula directly to the final hybrid cost. For a quantum parameter $\theta_k^{(\ell)}$ in block ℓ, $\frac{dC}{d\theta_k^{(\ell)}} = \sum_{j=0}^{3} \frac{\partial C}{\partial z_j^{(\ell)}} \frac{\partial z_j^{(\ell)}}{\partial \theta_k^{(\ell)}}$. (eq7).

The downstream sensitivity $\partial C / \partial z_j^{(\ell)}$ accumulates all paths from that feature to the final objective and is obtained by reverse-mode differentiation through later classical maps and quantum-node input Jacobians. For a Pauli-generated gate that appears once in the node, obtain independent feature estimates from positive and negative shifts of $\pi/2$. The corresponding node-level Jacobian estimate is the positive-shift estimate minus the negative-shift estimate, divided by two. The same rule is applied to $\partial z_j^{(\ell)} / \partial x_m^{(\ell)}$ when encoded input $x_m^{(\ell)}$ controls a single Pauli-generated rotation. If a shared parameter controls multiple gates, the derivative must instead be decomposed gate by gate or evaluated through a parameter-shift rule valid for the complete shared-parameter generator; the ordinary two-evaluation rule must not be assumed.

The derivatives of the trainable classical maps are evaluated exactly. Defining $\delta^{(\ell+1)} = \partial C / \partial x^{(\ell+1)}$, the classical-parameter gradients are $\frac{\partial C}{\partial W_\ell} = \delta^{(\ell+1)} \left(z^{(\ell)} \right)^T$, $\frac{\partial C}{\partial b_\ell} = \delta^{(\ell+1)}$.

The fixed shortcut gain γ has no trainable gradient in the primary factorial experiment. The exact classical derivatives are combined with estimated quantum-node Jacobians to construct the total gradient supplied to the optimizer. A direct parameter-shift rule may still be applied to a parameter in the final quantum node when the measured final objective is itself a supported expectation value, but it is not used as a shortcut for upstream parameters separated from the objective by nonlinear measured-feature re-encoding.

5.2.3 *SNR estimation and fair measurement budgets*

Estimator SNR is evaluated pointwise for the total derivative of the complete hybrid objective. For each fixed configuration, depth, initialization, and matched quantum-gate parameter, a finite-shot replicate first estimates the required forward feature vectors and then every node-level quantum Jacobian in Equation 7 from independent shifted shot batches. Exact classical Jacobians are used to assemble the replicate total-gradient estimate $\hat{g}_k$ through reverse-mode chain-rule propagation. Independent batches are used for distinct quantum Jacobian factors unless their covariance is explicitly estimated and incorporated.

The primary confirmatory SNR analysis is conducted at the matched initial parameter vectors, immediately after initialization and before the first optimizer update. This location permits the same quantum-gate parameters and initialization seeds to be compared across all eight configurations. The primary residual shortcut is already active through the fixed gain $\gamma = 1$, while the baseline uses $\gamma = 0$; therefore H3 and H4 do not compare architectures that are functionally identical at initialization. To assess whether gradient usability is preserved rather than only initialized favorably, a secondary longitudinal analysis evaluates SNR at prespecified checkpoints corresponding to 25%, 50%, 75%, and 100% of a fixed total optimization-evaluation budget. Because parameter trajectories diverge after the first update, checkpoint measurements are not treated as matched observations at identical parameter settings and are analyzed separately as within-configuration longitudinal diagnostics. Final convergence, energy, and fidelity remain distinct task-performance outcomes.

In ideal statevector simulation, exact feature values and node Jacobians are propagated through the same chain rule to obtain the reference total derivative g_k^{exact}. The primary finite-shot analysis re-estimates forward features within every replicate because those noisy features would be passed to later blocks in an end-to-end implementation. Nonlinear downstream re-encoding can make this estimator biased at finite shot count; the replicate mean therefore estimates $\mu_k = E_{\text{shots}}\left[\hat{g}_k\right]$, while variance and bias $Bias_k = \mu_k - g_k^{\text{exact}}$ are reported separately. The primary SNR^{est} uses $\left|\mu_k\right|$ in the numerator. When an exact statevector derivative or prespecified high-precision reference is available, SNR^{exact}, absolute bias,

and sign agreement between μ_k and g_k^{exact} are reported as complementary checks of estimator fidelity. A secondary conditional analysis holds forward features fixed at their exact statevector values in noiseless simulation, or at a prespecified high-precision reference on hardware, while resampling only node-level shifted measurements. This analysis isolates Jacobian-estimation noise from uncertainty introduced by repeatedly estimating and re-encoding intermediate features; the additional reference-shot cost must be included in resource accounting.

For every configuration, depth, initialization, matched quantum-gate parameter, and shot budget, $R = 30$ independent replicate gradients serve only as a minimum pilot starting value. Before confirmatory analysis, an independent pilot spanning representative configurations, depths, and shot budgets is used to estimate Monte Carlo uncertainty in the signed mean, sampling variance, and pointwise SNR^{est}. The final replicate count is the smallest prespecified increment above 30 for which the 95% bootstrap intervals for the signed mean and SNR satisfy stated absolute or relative half-width tolerances and change by less than a prespecified stability tolerance after the next increment in R. Pilot cells, tolerances, the increment rule, and the resulting R must be reported, and pilot data used to select R are excluded from confirmatory analysis.

The number of matched random initializations is selected through a separate simulation-based precision analysis rather than fixed at 50 by convention. Fifty initializations serve as the minimum candidate count. Using variance components and cross-configuration correlations estimated from an independent pilot, complete factorial datasets are simulated for candidate counts increasing in increments of 25. The confirmatory count is the smallest candidate for which the 90th percentile of the simulated 95% confidence-interval half-width is no greater than 0.20 for each primary estimator-SNR interaction coefficient β_{EL}, β_{ER}, and β_{LRd} and for the exact-signal interaction coefficient η_{EL} defined in Section 5.3. The selected count, pilot variance estimates, simulation procedure, and achieved interval-width distributions must be reported. Pilot initializations used to determine the count are excluded from the confirmatory dataset.

The same initialization seeds, block indices, and quantum-gate parameter indices are reused across configurations. Gradients of W_ℓ and b_ℓ are reported in a separate classical-parameter analysis and are not pooled with the matched quantum-gate gradients used for the primary factorial estimator-SNR comparison. On hardware, the signed replicate mean and sample variance estimate μ_k and its sampling variance. Because the plug-in absolute-mean SNR is upward-biased near a true zero signal, bootstrap intervals, the signed-mean interval, and bias relative to the best available reference must also be reported. An estimator is not described as resolved merely because its plug-in SNR^{est} exceeds one when the signed-mean interval includes zero, nor as faithful merely because it is resolved when bias or sign disagreement is material.

Each replicate receives the same total measurement budget B per matched total-gradient component and configuration. This

budget includes unshifted forward-feature estimates and every shifted circuit evaluation required by the chain-rule computational graph; it is not a separate budget for each node or observable. A prespecified allocation rule divides B across forward features, node-level parameter shifts, input-angle Jacobians, Pauli terms, and overlap-circuit settings. Equal allocation provides a transparent baseline, while allocation optimized from pilot variances may be reported as a sensitivity analysis. Reporting "1000 shots" without stating whether that number applies per circuit, observable, node-level derivative, replicate, or complete gradient would not define a comparable uncertainty scale. A prespecified sweep, such as $B \in \{250, 500, 1000, 2000\}$, tests whether interaction conclusions remain stable across sampling regimes. Total resource reporting must include R, every node-level shift, all observable groups, forward-feature measurements, initializations, checkpoints, and optimization evaluations.

Primary descriptive summaries are the median pointwise SNR^{est} of matched quantum-gate gradients, its interquartile range, the fraction with $SNR^{\text{est}} \geq 1$, the fraction whose signed-mean interval excludes zero, and the estimator-bias distribution. When a reference derivative is available, complementary summaries include median SNR^{exact}, absolute bias, and the fraction of matched components with the correct sign. Secondary measures include root-mean-square exact total-gradient magnitude, $Var_{\theta}\left[g_k^{\text{exact}}\right]$, classical-parameter gradient diagnostics, longitudinal SNR at prespecified checkpoints, full-energy convergence, global fidelity, variability across optimization seeds, entanglement entropy or purity, hardware-noise sensitivity, and total circuit evaluations. Ideal statevector, conditional finite-shot, end-to-end finite-shot, and hardware-noisy conditions are reported separately so that exact landscape structure, node-level sampling uncertainty, finite-shot feature propagation, and systematic hardware bias are not conflated. The results characterize finite-system trainability proxies and optimization behavior rather than definitive mitigation of asymptotic barren plateaus.

5.3 Testable hypotheses

The framework prespecifies four two-sided confirmatory interaction hypotheses and two exploratory performance questions. Because interaction is scale-dependent, exact-gradient signal and estimator SNR are analyzed in separate matched factorial models. Let E_c, L_c, and R_c indicate restricted entanglement, local cost, and residual connections for configuration c; let $\tilde{d}$ denote circuit depth centered and scaled to unit standard deviation; and let $s_b = \log_2(B_b)$ denote shot-budget scale.

For the exact-signal analysis at matched initial parameters, define $a_{icpd} = asinh\left(\left|g_{icpd}^{\text{exact}}\right|\right)$. A matched mixed-effects factorial model includes all intervention main effects, all pairwise products, the three-way product, centered depth, the lower-order intervention×depth terms, an initialization-level random intercept, and a matched-parameter random intercept nested within initialization. The restricted-entanglement×local-cost coefficient in this model is denoted η_{EL}. H1 tests $\eta_{EL} = 0$ two-sided. A negative estimate is interpreted as consistent with the prespecified sub-additive expectation, whereas a positive estimate is consistent with

super-additivity; neither sign is assumed by the test.

For estimator SNR, the primary inferential analysis uses a matched mixed-effects factorial regression on $asinh\left(SNR^{\text{est}}\right)$, which is defined at zero and is approximately logarithmic for large positive values. For initialization i, matched quantum-gate parameter p, configuration c, depth d, and budget b, the model is

$$y_{icpdb} = \beta_0 + \beta_E E_c + \beta_L L_c + \beta_R R_c + \beta_{EL} E_c L_c + \beta_{ER} E_c R_c + \beta_{LR} L_c R_c + \beta_{ELR} E_c L_c R_c + \beta_d \tilde{d} + \beta_B s_b + \beta_{Ed} E_c \tilde{d} + \beta_{Ld} L_c \tilde{d} + \beta_{Rd} R_c \tilde{d} + \beta_{LRd} L_c R_c \tilde{d} + u_i + v_{ip} + \varepsilon_{icpdb}.$$

Here, $y_{icpdb} = asinh\left(SNR^{\text{est}}_{icpdb}\right)$, u_i is an initialization-level random intercept, and v_{ip} is a random intercept for the matched parameter nested within initialization. Lower-order intervention×depth terms preserve model hierarchy for the depth-dependent residual× local-cost test. H2 tests $\beta_{EL} = 0$ at the mean depth, H3 tests $\beta_{ER} = 0$, H4 tests $\beta_{LRd} = 0$, and the three-way coefficient β_{ELR} remains exploratory. Additional depth- or budget-moderation terms not included in the prespecified primary models may be reported only as labeled sensitivity analyses rather than selected post hoc.

Depth moderation is prespecified as confirmatory only for the local-cost×residual interaction because Pathway C in Section 5.1 supplies an explicit literature-grounded reason for that pair's effect to vary with depth. H2 and H3 therefore estimate the restricted-entanglement×local-cost and restricted-entanglement×residual interactions at the centered mean depth. The confirmatory model omits $E_c L_c \tilde{d}$ and $E_c R_c \tilde{d}$ because no comparable directional or mechanistic depth-moderation hypothesis was formulated for those pairs and adding them would enlarge the confirmatory hypothesis family. A preregistered sensitivity model adds both terms while preserving all lower-order components. Those coefficients are reported descriptively and are not used for confirmatory claims; if their inclusion materially changes β_{EL} or β_{ER}, the corresponding primary effects are described explicitly as depth-averaged rather than generalized across the full sweep.

Confidence intervals are obtained through a nested matched bootstrap so that uncertainty in each pointwise SNR is propagated into the interaction coefficients. At the outer level, complete matched initializations are resampled, retaining all eight configurations, matched parameter indices, depths, and budgets for each selected initialization. Within every selected initialization–configuration–parameter–depth–budget cell, the R finite-shot replicate gradients are resampled; the signed mean, sampling variance, and SNR^{est} are recomputed; and the primary regression is refit. This procedure preserves cross-configuration matching while propagating both between-initialization variability and finite-R estimation uncertainty. Individual parameter observations are not resampled as though they were independent.

H1–H4 form a single confirmatory hypothesis family and all four tests are two-sided. Family-wise type-I error is controlled at $\alpha = 0.05$ with the Holm–Bonferroni procedure applied to the

four confirmatory *p*-values. Effect estimates and unadjusted 95% confidence intervals are reported for interpretation, but a confirmatory claim requires the corresponding Holm-adjusted *p*-value to meet the family-wise threshold. The three-way interaction, checkpoint analyses, and two performance questions are labeled exploratory and are not used to make multiplicity-adjusted confirmatory claims.

For descriptive fold-change interpretation, let M_A denote the root-mean-square pointwise SNR^{est} for configuration A and G_A the root-mean-square exact-gradient magnitude, each calculated over the same matched parameter points. RMS is used because both outcomes are magnitude-based and because it summarizes total squared resolvability or signal without cancellation between positive and negative gradients while retaining the original units. RMS gives greater weight to unusually large components; Equation 8 is therefore secondary and must be accompanied by the median, interquartile range, full matched distribution, and a median-based fold-change sensitivity analysis. For pair A, B, define, $I_{AB} = \frac{M_{AB} M_0}{M_A M_B}, J_{AB} = \frac{G_{AB} G_0}{G_A G_B}$. (eq 8).

Values below, equal to, or above one describe sub-additive, multiplicatively independent, or super-additive fold changes on the selected scale. These indices are secondary summaries: if an aggregate is numerically zero, the ratio is undefined and the factorial model remains primary. Additive contrasts on untransformed outcomes are reported as a sensitivity analysis so that conclusions do not depend on a single interaction scale.

5.3.1 *H1 (exact-gradient signal; two-sided test)*
Do restricted entanglement and the local TFIM-energy objective exhibit a nonzero interaction in exact-gradient magnitude at matched initial parameters? The prespecified hypotheses are $H_0: \eta_{EL} = 0$ and $H_A: \eta_{EL} \neq 0$. The mechanistic discussion in Section 5.1.1 supplies an interpretive expectation of a negative, sub-additive coefficient, but Equation 4 does not justify a one-sided test. A confidence interval entirely below zero supports sub-additivity on the prespecified $asinh(|g^{\text{exact}}|)$ scale; an interval entirely above zero supports super-additivity; and an interval containing zero does not establish either interaction direction.

5.3.2 *H2 (complete estimator SNR; two-sided test)*
Does restricted entanglement combined with the local cost produce a nonzero estimator-SNR interaction under matched measurement budgets? The prespecified hypotheses are $H_0: \beta_{EL} = 0$ and $H_A: \beta_{EL} \neq 0$; no directional sign is assigned. Estimator mean, finite-shot variance, and bias may preserve, offset, or reverse the signal-level interaction observed or expected under H1. Configuration [5] is compared with configurations [1], [2], and [3].

5.3.3 *H3 (complete estimator SNR; two-sided test)*
Does restricted entanglement combined with residual connections produce a nonzero estimator-SNR interaction under matched measurement budgets? The prespecified hypotheses are $H_0: \beta_{ER} = 0$ and $H_A: \beta_{ER} \neq 0$; no directional sign is assigned. Configuration [6] is compared with configurations [1], [2], and [4]. For H2 and H3, failure to reject a zero

interaction is not evidence of equivalence unless a practically negligible tolerance and corresponding equivalence test are prespecified.

5.3.4 *H4 (depth moderation; two-sided test)*
Does the residual×local-cost estimator-SNR interaction vary with circuit depth? The prespecified hypotheses are $H_0: \beta_{LRd} = 0$ and $H_A: \beta_{LRd} \neq 0$; no universal sign is assigned. Configuration [7] and the depth sweep identify the residual×local-cost interaction and its depth moderation.

Two performance questions remain exploratory: whether configuration [8] exceeds the best single intervention in SNR^{est},and whether any estimator-SNR advantage survives evaluation by final TFIM energy, global fidelity, and total circuit cost. The three-way interaction and longitudinal checkpoint analyses are likewise exploratory. These analyses characterize interaction direction, magnitude, depth dependence, and task trade-offs rather than confirm that combining strategies is beneficial.

5.4 Trade-offs and practical considerations
Each mitigation strategy introduces computational and representational trade-offs that must be included in evaluation. Restricted entanglement may improve gradient usability while reducing the set of representable states. A local cost may provide stronger local gradients while failing to distinguish globally different states with similar reduced marginals. Residual connections may improve optimization while adding classical parameters or feature-processing overhead. A configuration should therefore not be judged successful from estimator SNR alone; it must also reach competitive final energy and global fidelity under a comparable circuit-evaluation budget.
Measurement cost requires particular care. Global and local objectives may require different numbers of circuits or observable settings, so equal shots per setting would allocate unequal total resources. The primary comparison therefore uses an equal total budget per gradient component, with a sensitivity analysis that instead normalizes budget per observable. Hardware noise should be introduced through a specified model or measured device, and ideal statevector, finite-shot noiseless, and hardware-noisy conditions should be reported separately. This separation identifies whether an interaction arises from the exact landscape, finite sampling, or device error. The factorial experiment should be preregistered at the level of configurations, depth values, shot budgets, initialization distribution, SNR measurement location, checkpoint schedule, replicate and initialization selection rules, primary estimator-SNR summary, interaction model, and multiplicity procedure. Because the framework generates several outcomes, post hoc selection of a favorable parameter, depth, checkpoint, or metric could create a misleading appearance of complementarity. Reporting all eight configurations, all four confirmatory hypotheses, and confidence intervals for every pairwise interaction is necessary for falsifiability.

5.5 Limitations of few-qubit evaluation
In few-qubit systems, the exponential scaling that defines barren plateaus asymptotically generally cannot be established from a limited numerical sweep without additional theoretical support. The proposed framework addresses this

limitation by adopting pointwise estimator SNR under finite sampling as a primary operational metric while retaining landscape variance as a separate secondary diagnostic. The evaluation criteria are thus calibrated to quantities directly measurable in 2–10-qubit systems.

This calibration does not remove finite-size limitations. Four-qubit entropy measurements cannot establish area-law scaling; a favorable estimator-SNR distribution cannot prove the absence of a barren plateau at larger n; and successful convergence on the TFIM target does not guarantee the same interaction in classification or non-local optimization. The experiment should be interpreted as a controlled test of interaction mechanisms and gradient usability under stated finite-shot conditions. Claims about asymptotic barren-plateau mitigation, universal optimizer thresholds, or broad task generality should not be extrapolated from the proposed results.

The framework also considers only one implementation from each of three mitigation categories.

Other entangling structures, local observables, residual implementations, parameter initializations, or error-mitigation procedures may produce different interactions. A confirmed effect would therefore establish behavior for the tested design rather than a universal law for the category. In particular, the cost-function coefficient represents the contrast between normalized TFIM energy and global target-state infidelity, not measurement locality in isolation, and the residual coefficient represents the fixed-gain shortcut specified in Section 5.2. The broader contribution is methodological: demonstrating how complete factorial comparisons, pointwise estimator SNR, exact-gradient references, and bias diagnostics can distinguish signal, sampling, and interaction effects in few-qubit QNN studies.

6. Conclusion

Barren plateaus are among the most significant obstacles to the practical use of quantum neural networks on near-term NISQ hardware. This review examined the mechanisms that produce untrainable landscapes—including random entanglement, global cost functions, hardware noise, excessive depth, and highly expressive state distributions—and the strategies developed to address them. Across these approaches, the central design problem is not to maximize expressibility or gradient magnitude in isolation, but to preserve sufficient task-relevant information for optimization within realistic measurement and noise constraints.

The review's primary contribution is to reframe gradient-based trainability evaluation in the few-qubit regime around pointwise estimator SNR rather than asymptotic gradient-variance scaling alone. As established in Section 1.2.1, the reframing follows from three constraints. For gradient-based methods, practical progress is influenced by the usability of the gradient estimates rather than merely the mathematical existence of nonzero derivatives; finite-shot hybrid gradient estimation introduces configuration-dependent uncertainty; and a numerical sweep across 2–10 qubits generally characterizes finite-size concentration more directly than it determines an asymptotic decay class. Estimator SNR therefore measures the resolvability of the information supplied to the optimizer under a specified estimator and

budget, while exact-gradient references and bias diagnostics remain necessary for evaluating whether that information is faithful. This operational role does not make pointwise SNR a universal necessary condition for stochastic convergence.

The definition is intentionally pointwise. At a fixed parameter setting, the estimator mean μ_k is compared with the shot-induced standard deviation of $\hat{g}_k$; the resulting SNR^{est} values are then summarized across parameters and initializations. The exact derivative g_k^{exact} is retained to calculate estimator bias, sign agreement, and, where possible, SNR^{exact}. This preserves the distinction among exact-gradient landscape variance $Var_\theta\left[g_k^{\text{exact}}\right]$, estimator mean, and finite-shot variance $Var_{\text{shots}}\left[\hat{g}_k \mid \theta\right]$. In the proposed experiment, the primary confirmatory SNR analysis is performed at matched initial parameters before optimization begins, while prespecified checkpoint measurements provide a separate longitudinal diagnostic after trajectories diverge. None of these quantities alone guarantees convergence, and no universal gradient threshold follows from shot count without specifying the complete estimator.

The second contribution is a mechanistically motivated, fully testable framework for evaluating interactions among restricted entanglement, local costs, and residual connections. The framework does not claim that the strategies combine constructively. It converts uncertainty about their joint behavior into a complete 2^3 factorial experiment containing the baseline, all three single interventions, all three pairwise combinations, and the fully combined model. Including the local-cost×residual condition is essential because it permits the depth-dependent H4 interaction to be identified independently rather than inferred from a configuration containing all three components.

The strongest mechanistic expectation concerns restricted entanglement paired with the local TFIM-energy objective at the level of exact-gradient magnitude. Equation 4 shows how the entangling architecture alters reduced-state derivatives while the local objective selects and weights the components contributing to the measured gradient. Because the terms can reinforce or cancel, the exact-gradient interaction is tested two-sided; sub-additivity is used only to interpret a negative estimate. The identity does not determine finite-shot variance or the sign of the complete estimator-SNR interaction. Interactions involving residual connections remain non-directional or depth-dependent questions and are tested with an active fixed-gain shortcut at the matched initial parameter point.

The proposed experiment addresses several confounds that would otherwise prevent valid inference. Global and local objectives share the same target ground state, and final energy and global fidelity are reported for every configuration. Restricted entanglement is verified through entropy or purity rather than presented as evidence of area-law scaling from four qubits. Every matched quantum-gradient component receives the same total measurement budget across the complete chain-rule computational graph, and estimator variance is measured through independent replicate batches. The number of replicates and independent matched initializations is selected

through prespecified pilot-based precision rules. Pairwise interactions are evaluated through the matched mixed-effects factorial model in Section 5.3, while Equation 8 supplies secondary fold-change summaries. Nested bootstrap intervals propagate finite-R uncertainty and between-initialization variability, and Holm–Bonferroni adjustment controls family-wise error across H1–H4. Additive-scale and longitudinal checkpoint analyses remain labeled sensitivity or exploratory analyses.

Among the strategies surveyed, ansatz design occupies a central role. Hardware-efficient ansätze limit circuit depth and compilation overhead, reducing exposure to hardware error but potentially restricting expressibility. Problem-inspired ansätze align circuit structure with target symmetries or locality, preserving task-relevant correlations at the cost of generality. Local costs, initialization strategies, residual architectures, and error management can complement either approach, but their benefits remain conditional on the circuit, task, and measurement protocol. The review therefore supports a design heuristic rather than a universal architecture: use a problem-inspired construction when the task maps onto a circuit family with demonstrated trainability properties, and otherwise use a hardware-efficient design as a controlled baseline.

The two contributions are related but not co-equal. The SNR reframing is the principal conceptual contribution because it synthesizes an operational criterion suited to finite-shot few-qubit studies; SNR itself has already appeared in adaptive-shot and gradient-precision work. The interaction framework is a secondary contribution and research agenda enabled by that criterion. The review provides a literature-grounded reason to evaluate gradient usability through pointwise estimator SNR, a structural reason to test whether one strategy pair interacts at the exact-gradient-signal level, and an implementable design capable of separating all pairwise interactions. It does not demonstrate an interaction outcome empirically, determine the sign of a complete estimator-SNR interaction, prove the absence of barren plateaus, or establish a universal threshold for optimizer success.

In the few-qubit regime, where each parameter, gate, and circuit evaluation carries substantial cost, the distinction between theoretical concentration and operational gradient usability is especially important. A circuit can possess a nonzero but practically unresolvable gradient, while favorable estimator SNR can coexist with poor expressibility or an unsuitable objective. A successful mitigation strategy must preserve measurable gradients and complete the intended task under a realistic resource budget. The framework proposed here supplies a method for evaluating that combined requirement. All conclusions remain scoped to 2–10 qubits and to the architectures and tasks reviewed. Independent validation across larger systems, alternative objectives, and additional mitigation combinations is necessary before broader claims about asymptotic barren-plateau mitigation can be made.

Acknowledgments

The authors conducted this review independently and received no laboratory or field supervision or funding. Large language model assistants—including Claude Sonnet 4.6

(Anthropic, accessed February-July 2026) and Gemini 3.1 Pro (Google, accessed February-May 2026)—were used to support literature searches, clarify technical concepts during research, and provide feedback on draft prose. The authors independently reviewed, verified, and revised all technical claims, citations, arguments, and final written content; no AI tool independently generated the manuscript's substantive intellectual contribution.

Abbreviations

CNOT, controlled-NOT gate; CX, controlled-X gate; HEA, hardware-efficient ansatz; HVA, Hamiltonian variational ansatz; NISQ, noisy intermediate-scale quantum; QAOA, Quantum Approximate Optimization Algorithm; QCNN, quantum convolutional neural network; QML, quantum machine learning; QNN, quantum neural network; RMS, root-mean-square; SNR, signal-to-noise ratio; TFIM, transverse-field Ising model; VQA, variational quantum algorithm; VQE, Variational Quantum Eigensolver.

7. References

1. Peruzzo, A., McClean, J., Shadbolt, P., Yung, M. H., Zhou, X.-Q., Love, P. J., Aspuru-Guzik, A., & O'Brien, J. L. (2014). *A variational eigenvalue solver on a photonic quantum processor*. *Nature Communications, 5*, Article 4213. https://doi.org/10.1038/ncomms5213

2. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). *Quantum machine learning*. *Nature, 549*(7671), 195–202. https://doi.org/10.1038/nature23474

3. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R., & Neven, H. (2018). *Barren plateaus in quantum neural network training landscapes*. *Nature Communications, 9*, Article 4812. https://doi.org/10.1038/s41467-018-07090-4

4. Cerezo, M., Sone, A., Volkoff, T., Cincio, L., & Coles, P. J. (2021). *Cost function dependent barren plateaus in shallow parametrized quantum circuits*. *Nature Communications, 12*, Article 1791. https://doi.org/10.1038/s41467-021-21728-w

5. Wang, S., Fontana, E., Cerezo, M., Sharma, K., Sone, A., Cincio, L., & Coles, P. J. (2021). *Noise-induced barren plateaus in variational quantum algorithms*. *Nature Communications, 12*, Article 6961. https://doi.org/10.1038/s41467-021-27045-6

6. Ortiz Marrero, C., Kieferová, M., & Wiebe, N. (2021). *Entanglement-induced barren plateaus*. *PRX Quantum, 2*(4), Article 040316. https://doi.org/10.1103/PRXQuantum.2.040316

7. Ragone, M., Bakalov, B. N., Sauvage, F., Kemper, A. F., Ortiz Marrero, C., Larocca, M., & Cerezo, M. (2024). A Lie algebraic theory of barren plateaus for deep parameterized quantum circuits. Nature Communications, 15, Article 7172. https://doi.org/10.1038/s41467-024-49909-3

8. Arrasmith, A., Cerezo, M., Czarnik, P., Cincio, L., & Coles, P. J. (2021). *Effect of barren plateaus on gradient-free optimization*. *Quantum, 5*, Article 558. https://doi.org/10.22331/q-2021-10-05-558

9. Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., Fujii, K., McClean, J. R., Mitarai, K., Yuan, X., Cincio, L., & Coles, P. J. (2021). *Variational quantum algorithms*. *Nature Reviews Physics, 3*(9), 625–644. https://doi.org/10.1038/s42254-021-00348-9

10. Larocca, M., Thanasilp, S., Wang, S., Sharma, K., Biamonte, J., Coles, P. J., Cincio, L., McClean, J. R., Holmes, Z., & Cerezo, M. (2025). *Barren plateaus in variational quantum computing*. *Nature Reviews Physics, 7*(4), 174–189. https://doi.org/10.1038/s42254-025-00813-9

11. Tamiya, S., & Yamasaki, H. (2022). *Stochastic gradient line Bayesian optimization for efficient noise-robust optimization of parameterized quantum circuits*. *npj Quantum Information, 8*(1), Article 90. https://doi.org/10.1038/s41534-022-00592-6

12. Mitarai, K., Negoro, M., Kitagawa, M., & Fujii, K. (2018). *Quantum circuit learning*. *Physical Review A, 98*(3), 032309. https://doi.org/10.1103/PhysRevA.98.032309

13. Schuld, M., Bergholm, V., Gogolin, C., Izaac, J., & Killoran, N. (2019). *Evaluating analytic gradients on quantum hardware*. *Physical Review A, 99*(3), 032331. https://doi.org/10.1103/PhysRevA.99.032331

14. Sweke, R., Wilde, F., Meyer, J., Schuld, M., Fährmann, P. K., Meynard-Piganeau, B., & Eisert, J. (2020). *Stochastic gradient descent for hybrid quantum-classical optimization*. *Quantum, 4*, Article 314. https://doi.org/10.22331/q-2020-08-31-314

15. Kreplin, D. A., & Roth, M. (2024). *Reduction of finite sampling noise in quantum neural networks*. *Quantum, 8*, Article 1385. https://doi.org/10.22331/q-2024-06-25-1385

16. Chinzei, K., Yamano, S., Tran, Q. H., Endo, Y., & Oshima, H. (2025). *Trade-off between gradient measurement efficiency and expressivity in deep quantum neural networks*. *npj Quantum Information, 11*, Article 79. https://doi.org/10.1038/s41534-025-01036-7

17. Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). *Power-law distributions in empirical data*. *SIAM Review, 51*(4), 661–703. https://doi.org/10.1137/070710111

18. Harada, K. (2011). *Bayesian inference in the scaling analysis of critical phenomena*. *Physical Review E, 84*(5), 056704. https://doi.org/10.1103/PhysRevE.84.056704

19. Holmes, Z., Sharma, K., Cerezo, M., & Coles, P. J. (2022). *Connecting ansatz expressibility to gradient magnitudes and barren plateaus*. *PRX Quantum, 3*(1), 010313. https://doi.org/10.1103/PRXQuantum.3.010313

20. Kashif, M., & Al-Kuwari, S. (2024). *ResQNets: A residual approach for mitigating barren plateaus in quantum neural networks*. *EPJ Quantum Technology, 11*, Article 4. https://doi.org/10.1140/epjqt/s40507-023-00216-8

21. Patti, T. L., Najafi, K., Gao, X., & Yelin, S. F. (2021). *Entanglement devised barren plateau mitigation*. *Physical Review Research, 3*(3), 033090. https://doi.org/10.1103/PhysRevResearch.3.033090

22. Fontana, E., Herman, D., Chakrabarti, S., Kumar, N., Yalovetzky, R., Heredge, J., Sureshbabu, S. H. and Pistoia, M. (2024). *Characterizing barren plateaus in quantum ansätze with the adjoint representation*. *Nature Communications, 15*, Article 7171. https://doi.org/10.1038/s41467-024-49910-w

23. Larocca, M., Czarnik, P., Sharma, K., Muraleedharan, G., Coles, P. J., & Cerezo, M. (2022). *Diagnosing barren plateaus with tools from quantum optimal control*. *Quantum, 6*, Article 824. https://doi.org/10.22331/q-2022-09-29-824

24. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

25. Grant, E., Wossnig, L., Ostaszewski, M., & Benedetti, M. (2019). *An initialization strategy for addressing barren plateaus in parametrized quantum circuits*. *Quantum, 3*, Article 214. https://doi.org/10.22331/q-2019-12-09-214

26. Liu, H. Y., Sun, T. P., Wu, Y. C., Han, Y. J., & Guo, G. P. (2023). *Mitigating barren plateaus with transfer-learning-inspired parameter initialisations*. *New Journal of Physics, 25*, 013039. https://doi.org/10.1088/1367-2630/acb58e

27. Kulshrestha, A., & Safro, I. (2022). *BEINIT: Avoiding barren plateaus in variational quantum algorithms*. In *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)* (pp. 197–203). IEEE. https://doi.org/10.1109/QCE53715.2022.00039

28. Cervero Martín, E., Plekhanov, K., & Lubasch, M. (2023). *Barren plateaus in quantum tensor network optimization*. *Quantum, 7*, Article 974. https://doi.org/10.22331/q-2023-04-13-974

29. Pesah, A., Cerezo, M., Wang, S., Volkoff, T., Sornborger, A. T., & Coles, P. J. (2021). *Absence of barren plateaus in quantum convolutional neural networks*. *Physical Review X, 11*(4), 041011. https://doi.org/10.1103/PhysRevX.11.041011

30. Cong, I., Choi, S., & Lukin, M. D. (2019). *Quantum convolutional neural networks*. *Nature Physics, 15*(12), 1273–1278. https://doi.org/10.1038/s41567-019-0648-8

31. Li, Y., & Benjamin, S. C. (2017). *Efficient variational quantum simulator incorporating active error minimization*. *Physical Review X, 7*(2), 021050. https://doi.org/10.1103/PhysRevX.7.021050

32. Temme, K., Bravyi, S., & Gambetta, J. M. (2017). *Error mitigation for short-depth quantum circuits*. *Physical Review Letters, 119*(18), 180509. https://doi.org/10.1103/PhysRevLett.119.180509

33. Endo, S., Cai, Z., Benjamin, S. C., & Yuan, X. (2021). *Hybrid quantum-classical algorithms and quantum error mitigation*. *Journal of the Physical Society of Japan, 90*(3), 032001. https://doi.org/10.7566/JPSJ.90.032001

34. Kandala, A., Mezzacapo, A., Temme, K., Takita, M., Brink, M., Chow, J. M., *et al.* (2017). *Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets*. *Nature, 549*(7671), 242–246. https://doi.org/10.1038/nature23879

35. Park, C. Y., & Killoran, N. (2024). *Hamiltonian variational ansatz without barren plateaus*. *Quantum, 8*, Article 1239. https://doi.org/10.22331/q-2024-02-01-1239

36. Du, Y., Hsieh, M.-H., Liu, T., & Tao, D. (2020). *The expressive power of parameterized quantum circuits*. *Physical Review Research, 2*(3), 033125. https://doi.org/10.1103/PhysRevResearch.2.033125

37. Kobayashi, M., Nakaji, K., & Yamamoto, N. (2022). *Overfitting in quantum machine learning and entangling dropout*. *Quantum Machine Intelligence, 4*, Article 30. https://doi.org/10.1007/s42484-022-00087-9

38. Tüysüz, C., Clemente, G., Crippa, A., Hartung, T., Kühn, S., & Jansen, K. (2023). *Classical splitting of parametrized quantum circuits*. *Quantum Machine Intelligence, 5*(2), Article 34. https://doi.org/10.1007/s42484-023-00118-z