IDENTITY-PAIRED PROGRESSIVE DEPTH TRAINING: WHEN TRAINABILITY PERSISTS BEYOND EXPRESSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Variational Quantum Algorithms (VQAs) are a leading paradigm for near-term quantum computing, yet their training suffers from sensitivity to circuit depth, initialization, and landscape pathologies such as barren plateaus. We study *progressive depth training* (PDT) a layerwise curriculum that trains a shallow circuit before appending new layers and identify a fundamental obstacle: fixed entangling gates (CNOTs) in hardware-efficient ansätze cause *initialization shock*, an energy spike when new layers are added. We propose *identity-paired progressive depth training* (IP-PDT), which appends forward/inverse block pairs each consisting of a standard rotation+CNOT block followed by its reverse that compose to the identity at initialization. Because the adjacent CNOT rings cancel, the effective circuit retains only *a single entangling layer* surrounded by *overparameterized local rotations*. We prove a simple *Reachable Set Saturation Theorem*: under this construction the variational manifold expands exactly once (when post-entangler rotations are first introduced) and then *saturates*; all subsequent depth increases provide pure overparameterization of single-qubit unitaries. Despite this saturation, progressive addition of rotation parameters can continue to improve optimization outcomes a phenomenon we term *trainability beyond expressibility*. We formalize IP-PDT as a continuation method on nested manifolds, prove monotone energy guarantees under an acceptance rule, and connect energy error to ground-state fidelity through spectral-gap inequalities. A detailed resource analysis shows that IP-PDT achieves lower total gate cost than both baselines by eliminating most CNOT gates. Experiments span six-qubit Transverse-Field Ising, Tilted Ising, and Random Ising Hamiltonians, a broader nine-Hamiltonian benchmark, and system sizes up to $n = 16$ qubits. IP-PDT matches or outperforms full-depth baselines on Hamiltonians whose ground states are well-approximated by the single-entangler reachable set, with particularly strong gains in limited-budget regimes. At $n = 16$ the hardware-efficient baseline degrades sharply while the single-entangler methods remain trainable.

1 INTRODUCTION

Variational quantum algorithms (VQAs) turn limited-depth quantum evolution into a classical optimization problem (Peruzzo et al., 2014; Cerezo et al., 2021a): minimize the energy $E(\boldsymbol{\theta}) = \langle \boldsymbol{\psi}(\boldsymbol{\theta}) | \mathbf{H} | \boldsymbol{\psi}(\boldsymbol{\theta}) \rangle$ of a state prepared by a parameterized quantum circuit (PQC). As depth grows, *expressibility* improves more states become reachable while *trainability* can degrade through vanishing gradients (McClean et al., 2018), rugged landscapes (Anschuetz & Kiani, 2022), and sensitivity to initialization.

A natural remedy is to treat depth as a *curriculum*: optimize a shallow circuit, then grow. This *progressive depth training* (PDT) strategy underlies ADAPT-VQE (Grimsley et al., 2019), layerwise training (Skolik et al., 2021), and initialization transfer (Grant et al., 2019). A practical obstacle remains under-addressed. In hardware-efficient ansätze (HEA) (Kandala et al., 2017) each layer contains fixed entangling gates (CNOTs) that have no parameter setting acting as the identity, so every depth expansion irreversibly commits to additional entanglement. Appending a layer often

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

disrupts the trained state, producing an *initialization shock*—an energy spike that can erase the progress of prior stages.

Our approach. We propose *identity-paired progressive depth training* (IP-PDT), in which each expansion appends a *pair* of blocks one standard (rotations → CNOT ring) and one reverse (reversed CNOT ring → inverse rotations) that compose to the identity at initialization. The adjacent CNOT rings cancel exactly ($\mathbf{U}_{\text{ent}}\mathbf{U}_{\text{ent}}^{-1} = \mathbf{I}$), so the effective circuit contains *a single entangling layer* surrounded by *overparameterized local rotations*: every expansion beyond the first adds no entanglement at all. Yet, as we show, progressive training on these redundant parameters continues to yield consistent optimization gains. This challenges the intuition that deeper entanglement is what makes VQAs work, and points to a regime where *trainability gains persist beyond expressibility gains*. Our contributions:

1. *IP-PDT* (Section 2): a growth rule that eliminates initialization shock and collapses the trained circuit to a single entangling layer, with a gate-count analysis showing the lowest total training cost among the protocols we compare.
2. *Reachable Set Saturation* (Theorem 3.2): the variational manifold of IP-PDT circuits expands *exactly once* and then saturates; further expansions are pure overparameterization of $\mathrm{SU}(2)$ rotations.
3. *Continuation and monotonicity* (Section 3): we cast IP-PDT as homotopy continuation on nested manifolds, track the shallow-stage minimizer under expansion via the implicit function theorem, prove unconditional monotone energy descent under an acceptance rule, and show that gradient descent on the previous-stage parameters contracts geometrically within the inherited basin (joint dynamics are left open). A spectral-gap sandwich turns energy error into a ground-state fidelity certificate.
4. *Experiments* (Section 4): across nine Hamiltonians at $n = 6$ and $n = 16$ qubits and against five VQE baselines, IP-PDT matches or beats full-depth HEA with $5\times$ fewer CNOTs, with the largest gains at low budgets; ablations show the progressive curriculum, not the jitter, drives the gain.

Related work. *Progressive and layerwise VQE.* ADAPT-VQE (Grimsley et al., 2019) grows circuits by greedily selecting operators from a pool, at considerable cost and, to our knowledge, without closed-form convergence guarantees; Skolik et al. (2021) train layerwise with zero-initialized new parameters but do not address CNOT-induced shock. Closest in spirit, Žunkovič et al. (2026) warm-start VQE along an adiabatic path of Hamiltonians continuation in Hamiltonian space, complementary to our depth-space curriculum. Classically, Szlendak et al. (2024) cast progressive training as randomized coordinate descent. We instead exploit circuit structure to obtain *lossless* depth growth with provable monotonicity. *Identity-block initialization.* Grant et al. (2019) initialize individual gates near the identity to mitigate barren plateaus; we pair entire HEA blocks (including CNOTs) and exploit the resulting cancellation, which *removes* entangling gates rather than initializing them. *Overparameterization.* Larocca et al. (2023) locate the onset of overparameterization via the dynamical Lie algebra (DLA); Holmes et al. (2022) link expressibility to gradient magnitudes. Our saturation result concerns a different invariant: the reachable *set* of the single-entangler ansatz stops growing at $D = 2$, with only $O(n)$ parameters per layer rather than the $4^n - 1$ dimensions of $\mathfrak{su}(2^n)$ a deep HEA can explore. Computing the SE DLA exactly is left open.

2 Identity-Paired Progressive Depth Training

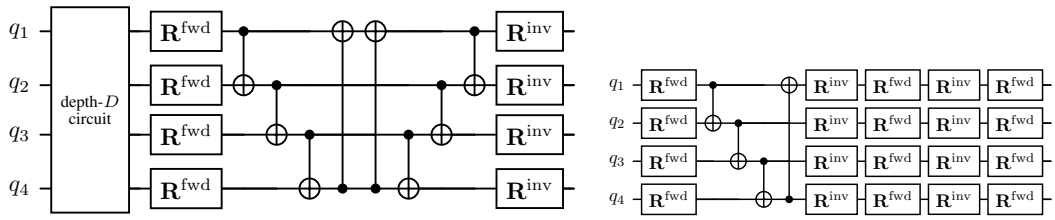
Setup. We work on n qubits, $\mathcal{H} \cong (\mathbb{C}^2)^{\otimes n}$, $d = 2^n$, with a Hamiltonian $\mathbf{H} = \sum_i \lambda_i |\mathbf{v}_i\rangle\langle\mathbf{v}_i|$, $\lambda_1 \le \cdots \le \lambda_d$, and spectral gap $\Delta = \lambda_2 - \lambda_1$. A PQC $\mathbf{U}(\boldsymbol{\theta})$ prepares $|\psi(\boldsymbol{\theta})\rangle = \mathbf{U}(\boldsymbol{\theta})|0^n\rangle$ and we minimize

$$E(\boldsymbol{\theta}) = \langle 0^n|\mathbf{U}(\boldsymbol{\theta})^\dagger \mathbf{H}\mathbf{U}(\boldsymbol{\theta})|0^n\rangle \ \ge \lambda_1. \tag{1}$$

Pauli rotations are $\mathbf{R}_\sigma(\theta) = \exp(-i\frac{\theta}{2}\sigma)$. For $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ define the forward single-qubit rotation and its inverse,

$$\mathbf{R}^{\text{fwd}}(\alpha,\beta,\gamma) = \mathbf{R}_Y(\alpha)\mathbf{R}_Z(\beta)\mathbf{R}_X(\gamma), \quad \mathbf{R}^{\text{inv}}(\alpha,\beta,\gamma) = \mathbf{R}_X(-\gamma)\mathbf{R}_Z(-\beta)\mathbf{R}_Y(-\alpha) = [\mathbf{R}^{\text{fwd}}]^\dagger, \tag{2}$$

so that $\mathbf{R}^{\text{fwd}}$ ranges over all of $\mathrm{SU}(2)$. For a parameter block $\boldsymbol{\theta}^{(d)} \in \mathbb{R}^{3n}$ the *forward rotation layer* is $\mathcal{R}^{\text{fwd}}(\boldsymbol{\theta}^{(d)}) = \bigotimes_{q=1}^n \mathbf{R}^{\text{fwd}}(\alpha_q^{(d)}, \beta_q^{(d)}, \gamma_q^{(d)})$ and $\mathcal{R}^{\text{inv}} = [\mathcal{R}^{\text{fwd}}]^\dagger$. The entangler is a CNOT ring,



(a) Identity-pair expansion $D \to D{+}2$.

(b) Effective SE circuit at $D = 5$.

Figure 1: (a) Each expansion appends a forward block and its reverse; the adjacent CNOT rings cancel. (b) After cancellation only the layer-0 ring survives; all later layers are local rotations ($n = 4$).

$\mathbf{U}_{\mathrm{ent}} = \prod_{q=1}^{n} \mathrm{CNOT}_{q\to(q \bmod n)+1}$, and its reverse $\mathbf{U}_{\mathrm{ent}}^{-1}$ applies the same gates in the opposite order, so $\mathbf{U}_{\mathrm{ent}}\mathbf{U}_{\mathrm{ent}}^{-1} = \mathbf{I}_{2^n}$. Full background and a notation table are in Appendix A.

Hardware-efficient ansatz. The standard HEA stacks D blocks $\mathbf{B}_d(\boldsymbol{\theta}^{(d)}) = \mathbf{U}_{\mathrm{ent}}\mathcal{R}^{\mathrm{fwd}}(\boldsymbol{\theta}^{(d)})$:

$$\mathbf{U}_D^{\mathrm{HEA}}(\boldsymbol{\theta}) = \prod_{d=D-1}^{0} \mathbf{B}_d(\boldsymbol{\theta}^{(d)}), \qquad \boldsymbol{\theta} \in \mathbb{R}^{3nD}, \tag{3}$$

time-ordered right to left. Because $\mathbf{U}_{\mathrm{ent}}$ is fixed, adding a block *always* injects a new entangling layer, which generically disrupts the state prepared so far. This is the source of initialization shock in naïve PDT.

Identity-pair expansion. To grow without disruption we append *two* blocks: a forward block and its reverse $\mathbf{B}^{\mathrm{rev}}(\boldsymbol{\theta}) = \mathcal{R}^{\mathrm{inv}}(\boldsymbol{\theta})\mathbf{U}_{\mathrm{ent}}^{-1}$.

Proposition 2.1 (Identity-pair cancellation)**.** *For any $\boldsymbol{\theta} \in \mathbb{R}^{3n}$,*

$$\mathbf{B}^{\mathrm{rev}}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta}) = \mathcal{R}^{\mathrm{inv}}(\boldsymbol{\theta})\,[\mathbf{U}_{\mathrm{ent}}^{-1}\mathbf{U}_{\mathrm{ent}}]\,\mathcal{R}^{\mathrm{fwd}}(\boldsymbol{\theta}) = \mathbf{I}_{2^n}. \tag{4}$$

The two CNOT rings are adjacent, so they cancel exactly and *need not be physically implemented*; the appended pair reduces to $\mathcal{R}^{\mathrm{inv}}(\boldsymbol{\theta}^{(d+1)})\mathcal{R}^{\mathrm{fwd}}(\boldsymbol{\theta}^{(d)})$, purely local gates (Figure 1). Iterating, the IP-PDT circuit at "depth" D (D rotation layers) contains the CNOT ring only at layer 0:

$$\mathbf{U}_D^{\mathrm{SE}}(\boldsymbol{\theta}) = \Big(\prod_{k=D-1}^{1} \mathbf{W}_k(\boldsymbol{\theta}^{(k)})\Big)\mathbf{U}_{\mathrm{ent}}\,\mathcal{R}^{\mathrm{fwd}}(\boldsymbol{\theta}^{(0)}), \qquad \mathbf{W}_k = \begin{cases}\mathcal{R}^{\mathrm{inv}}(\boldsymbol{\theta}^{(k)}), & k \text{ odd},\\ \mathcal{R}^{\mathrm{fwd}}(\boldsymbol{\theta}^{(k)}), & k \geq 2 \text{ even}.\end{cases} \tag{5}$$

We call (5) the *single-entangler* (SE) circuit. Layers k (odd) and $k{+}1$ form an *identity pair* if $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k+1)}$, since then $\mathbf{W}_{k+1}\mathbf{W}_k = \mathcal{R}^{\mathrm{fwd}}\mathcal{R}^{\mathrm{inv}} = \mathbf{I}_{2^n}$ (Proposition B.2, Appendix B).

Identity-paired initialization and jitter. Expanding from depth D to $D{+}2$, we append layers D and $D{+}1$ as an identity pair copied from a reference vector $\boldsymbol{\theta}_{\mathrm{ref}}$, so that

$$E_{D+2}\big(\hat{\boldsymbol{\theta}}^{(0:D-1)}, \boldsymbol{\theta}_{\mathrm{ref}}, \boldsymbol{\theta}_{\mathrm{ref}}\big) = E_D\big(\hat{\boldsymbol{\theta}}^{(0:D-1)}\big) \tag{6}$$

for the previous-stage optimum $\hat{\boldsymbol{\theta}}^{(0:D-1)}$ (*zero-shock expansion*). To break the symmetry at the identity we perturb the new parameters,

$$\boldsymbol{\theta}^{(D)} = \boldsymbol{\theta}_{\mathrm{ref}} + \boldsymbol{\xi}^{(1)}, \quad \boldsymbol{\theta}^{(D+1)} = \boldsymbol{\theta}_{\mathrm{ref}} + \boldsymbol{\xi}^{(2)}, \quad \boldsymbol{\xi}^{(i)} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{3n}),\ \sigma \ll 1, \tag{7}$$

which hands the optimizer a descent direction while keeping the energy within $O(\sigma)$ of its pre-expansion value (Section 3). Algorithm 1 states the procedure for a stage schedule $\mathcal{D} = \{D_1, \ldots, D_M\}$; the optional acceptance step is the rule under which Theorem 3.4 holds.

Training protocols compared. All experiments compare three protocols that end with the same parameter count ($3nD_\star$, $D_\star = 5$) and spend the same budget ($3T$ gradient steps), seeding the first block from the same random draw: (a) *Baseline*: full-depth HEA (3) trained from scratch; (b) *Naïve PDT*: HEA grown $1 \to 3 \to 5$, each appended block carrying a fresh CNOT ring (copy-initialized

Algorithm 1 Identity-Paired Progressive Depth Training (IP-PDT)

Require: Stage depths $\mathcal{D} = \{D_1, \ldots, D_M\}$; steps per stage T; jitter σ; step size η.
Sample $\boldsymbol{\theta}_{\mathrm{ref}} \sim \mathrm{Unif}([0, 2\pi]^{3n})$; initialize $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_{\mathrm{ref}}$ at depth D_1.
for $m = 1$ to M **do**
    $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}$; run T steps $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla E_{D_m}(\boldsymbol{\theta})$.
    Accept (optional safeguard, see Theorem 3.4): if $E_{D_m}(\boldsymbol{\theta}) > E_{D_m}(\hat{\boldsymbol{\theta}})$, set $\boldsymbol{\theta} \leftarrow \hat{\boldsymbol{\theta}}$.
    if $m < M$ **then**
        Expand: append $(D_{m+1} - D_m)/2$ identity pairs with jitter (7).
    end if
end for
return $\boldsymbol{\theta}$.

from the shared first block unless stated otherwise); (c) *IP-PDT*: the SE circuit (5) grown $1 \to 3 \to 5$ by identity pairs. Since parameter count and budget are fixed, any difference in final energy comes from the circuit architecture and the growth rule (Appendix B.4 tabulates the three).

Resource estimation. Counting steps hides that a step through a deeper HEA does more work. At depth d the HEA uses $3nd$ rotations and nd CNOTs ($4nd$ gates); the SE circuit uses the same $3nd$ rotations but only n CNOTs *regardless of depth*, $n(3d + 1)$ gates. Summing per-step gate cost over the stages $\{1, 3, 5\}$ with T steps each, $C^{\mathrm{base}} = 60nT$, $C^{\mathrm{naive}} = 36nT$, $C^{\mathrm{IP}} = 30nT$ (Appendix B.5). On hardware, where two-qubit gates dominate error and wall-clock, the fivefold CNOT reduction is the more consequential saving.

3 Theoretical Analysis

Our experiments show that progressive training keeps helping after the circuit can no longer represent new states. The analysis answers four questions in sequence: *(i)* what can the SE circuit represent, and when does that stop growing? *(ii)* why does growing the circuit not destroy the solution already found? *(iii)* can we guarantee training never moves backwards? *(iv)* does low energy certify the ground state? Together: the single-entangler construction freezes expressibility early (i), so added depth is spent safely (ii) and monotonically (iii) on optimization, with a certificate tying energy to fidelity (iv). All proofs are in Appendices C, E and F.

3.1 Reachable set saturation

Let $\mathcal{M}_D = \{\mathbf{U}_D^{\mathrm{SE}}(\boldsymbol{\theta})|0^n\rangle : \boldsymbol{\theta} \in \mathbb{R}^{3nD}\}$. Because every post-entangler layer is a product of single-qubit unitaries, the layers on each wire compose into one $\mathrm{SU}(2)$ element:

Lemma 3.1 (State factorization)**.** *For any $D \geq 1$, every $|\psi\rangle \in \mathcal{M}_D$ can be written as*

$$|\psi\rangle = \big(\textstyle\bigotimes_q \mathbf{V}_q\big)\mathbf{U}_{\mathrm{ent}}\big(\textstyle\bigotimes_q r_q\big)|0^n\rangle \tag{8}$$

with $r_q, \mathbf{V}_q \in \mathrm{SU}(2)$, where $\mathbf{V}_q$ is the composition of all post-entangler rotations on qubit q.

Theorem 3.2 (Reachable Set Saturation)**.** *For the single-entangler family* (5)*: (a) $\mathcal{M}_1 \subseteq \mathcal{M}_2$; (b) $\mathcal{M}_D = \mathcal{M}_2$ for all $D \geq 2$; (c) if $\mathbf{U}_{\mathrm{ent}}$ does not normalize the local-unitary group $\mathcal{L} = \{\bigotimes_q \mathbf{V}_q : \mathbf{V}_q \in \mathrm{SU}(2)\}$ (as the CNOT ring does not), then $\mathcal{M}_1 \subsetneq \mathcal{M}_2$.*

Proof idea. (b) follows from Lemma 3.1: for $D \geq 2$ the post-entangler factor ranges over all of $\mathrm{SU}(2)^{\otimes n}$ already at $D = 2$, and extra layers only re-parameterize the same product. (c) uses that conjugating a local $\mathbf{R}_X$ by a CNOT yields an entangling XX rotation, which applied to a generic product state leaves $\mathcal{M}_1$; a SWAP, which *does* normalize $\mathcal{L}$, would give $\mathcal{M}_1 = \mathcal{M}_2$, so non-normalization is the correct hypothesis. □

The theorem turns the paper's central question—why does training keep improving?—into a question about *optimization*: for $D \geq 2$ the SE ansatz adds $O(n)$ parameters per layer but no new states.

3.2 CONTINUATION UNDER IDENTITY EXPANSION

Identity pairing gives the layered family an *identity-embedding property* (IEP): for every D there is a map $\iota_D : \mathbb{R}^{3nD} \to \mathbb{R}^{3n(D+2)}$ with $\mathbf{U}^{\mathrm{SE}}_{D+2}(\iota_D(\boldsymbol{\theta})) = \mathbf{U}^{\mathrm{SE}}_D(\boldsymbol{\theta})$ (Definition C.1). Hence $\mathcal{M}_D \subseteq \mathcal{M}_{D+2}$ and the optimal energies are nested, $E^\star_{D+2} \le E^\star_D$ (Proposition C.2 and Corollary C.3). To track a *particular* minimizer rather than the optimum, write the two new layers as $\exp(-i\mathbf{G}(\boldsymbol{\phi}))$ with Hermitian generator $\mathbf{G}(\boldsymbol{\phi}) = \sum_{j=1}^{6n} \phi_j \mathbf{G}_j + \mathbf{R}(\boldsymbol{\phi})$, $\|\mathbf{G}_j\| = \frac{1}{2}$, $\|\mathbf{R}(\boldsymbol{\phi})\| \le C_R\|\boldsymbol{\phi}\|^2$, and define the homotopy energy

$$F(\boldsymbol{\theta}, \boldsymbol{\phi}; \lambda) = \big\langle 0^n \big| \big[\mathbf{U}^{(\lambda)}_{D+2}\big]^\dagger \mathbf{H}\, \mathbf{U}^{(\lambda)}_{D+2} \big| 0^n \big\rangle, \qquad \mathbf{U}^{(\lambda)}_{D+2}(\boldsymbol{\theta}, \boldsymbol{\phi}) = e^{-i\lambda \mathbf{G}(\boldsymbol{\phi})}\, \mathbf{U}^{\mathrm{SE}}_D(\boldsymbol{\theta}), \tag{9}$$

so $F(\cdot; 0) = E_D$ and $F(\cdot; 1) = E_{D+2}$ (Appendix F.2).

Theorem 3.3 (Local continuation)**.** *Let $\boldsymbol{\theta}^\star_D$ be a local minimizer of E_D with $\nabla^2 E_D(\boldsymbol{\theta}^\star_D) \succeq \mu\mathbf{I}$, $\mu > 0$ (Assumption C.4); let F be C^2 near $(\boldsymbol{\theta}^\star_D, \boldsymbol{\phi}^\star, 0)$, $\boldsymbol{\phi}^\star \neq \mathbf{0}$, and $\partial_\lambda \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^\star_D, \boldsymbol{\phi}^\star; \lambda)|_{\lambda=0} \neq \mathbf{0}$. Then there exist $\lambda_{\max} \in (0, 1]$ and a unique C^1 curve $\boldsymbol{\theta} : [0, \lambda_{\max}] \to \mathbb{R}^{3nD}$ with $\boldsymbol{\theta}(0) = \boldsymbol{\theta}^\star_D$, $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}(\lambda), \boldsymbol{\phi}^\star; \lambda) = \mathbf{0}$, with initial velocity*

$$\boldsymbol{\theta}'(0) = -[\nabla^2 E_D(\boldsymbol{\theta}^\star_D)]^{-1}\, \partial_\lambda \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^\star_D, \boldsymbol{\phi}^\star; \lambda)|_{\lambda=0} \neq \mathbf{0}. \tag{10}$$

The hypothesis $\boldsymbol{\phi}^\star \neq \mathbf{0}$ is supplied by the jitter, and the coupling condition holds generically (its failure set has measure zero absent a Pauli-commutation symmetry of $\mathbf{H}$; Remark F.1). The tracked minimum moves at speed $O(\|\boldsymbol{\phi}^\star\|/\mu) = O(\sigma/\mu)$: an identity-pair expansion *deforms* the landscape already solved rather than dropping the optimizer into a fresh one. Quantitatively, the mean-value inequality applied to $t \mapsto e^{it\mathbf{G}}\mathbf{H}e^{-it\mathbf{G}}$ gives, with no series truncation,

$$|E_{D+2}(\boldsymbol{\theta}, \boldsymbol{\phi}) - E_D(\boldsymbol{\theta})| \le 2\|\mathbf{H}\|\, \|\mathbf{G}(\boldsymbol{\phi})\| \le 2\|\mathbf{H}\| \Big(\tfrac{1}{2} \textstyle\sum_j |\phi_j| + C_R\|\boldsymbol{\phi}\|^2 \Big), \tag{11}$$

so under (7) $\mathbb{E}|E_{D+2} - E_D| \le 6n\sigma\|\mathbf{H}\|\sqrt{2/\pi} + 12nC_R\sigma^2\|\mathbf{H}\|$ (Appendix E.2). The statement is deliberately local: it says a good minimum *persists* under expansion, not that gradient descent *finds* it. The next two results close that gap from two directions.

3.3 MONOTONICITY AND BASIN PRESERVATION

Theorem 3.4 (Monotone non-increase of produced energies)**.** *Under the acceptance rule of Algorithm 1 (accept if the energy improved, else fall back to the embedded parameters $\iota_D(\hat{\boldsymbol{\theta}}^{(D)})$), for all stages D:*

$$E_{D+2}(\hat{\boldsymbol{\theta}}^{(D+2)}) \le E_D(\hat{\boldsymbol{\theta}}^{(D)}). \tag{12}$$

The guarantee is unconditional—no assumption on landscape, Hamiltonian, or optimizer—because the IEP makes the fallback lossless; naïve PDT has no identity embedding and cannot implement it (Appendix F.4). For a *rate*, we localize.

Theorem 3.5 (Basin preservation for old parameters)**.** *Suppose E_D is L_D-smooth and μ_D-strongly convex on a bounded convex neighbourhood $\mathcal{N}_D$ of $\boldsymbol{\theta}^\star_D$ (Assumption C.7), fix the new-layer parameters at $\boldsymbol{\phi}^\star$ with $\|\boldsymbol{\phi}^\star\| = O(\sigma)$, and let $\boldsymbol{\theta}^\star_{D+2}$ minimize $E_{D+2}(\cdot, \boldsymbol{\phi}^\star)$ on $\mathcal{N}_D$. If the warm start $\boldsymbol{\theta}^{(0)}$ satisfies $\overline{B}(\boldsymbol{\theta}^\star_{D+2}, \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star_{D+2}\|) \subseteq \mathcal{N}_D$ and $\sigma \le \min(\sigma_0, \sigma_1)$ with $\sigma_0 = \mu_D/(2c)$, $c = O(n^{3/2}D\|\mathbf{H}\|)$ the Hessian-perturbation constant and σ_1 an interiority threshold (Appendix F.6), then $\mu_{D+2} \ge \mu_D/2$, $L_{D+2} \le L_D + c\sigma$, and for $\eta \in (0, 1/L_{D+2}]$ gradient descent on $\boldsymbol{\theta}$ stays in $\mathcal{N}_D$ with*

$$E_{D+2}(\boldsymbol{\theta}^+, \boldsymbol{\phi}^\star) - E_{D+2}(\boldsymbol{\theta}^\star_{D+2}, \boldsymbol{\phi}^\star) \le (1 - \eta\mu_{D+2}) \big(E_{D+2}(\boldsymbol{\theta}, \boldsymbol{\phi}^\star) - E_{D+2}(\boldsymbol{\theta}^\star_{D+2}, \boldsymbol{\phi}^\star) \big). \tag{13}$$

Scope. The theorem concerns descent on the *previous-stage* parameters with the appended layers held fixed; joint dynamics are left open (Appendix C.5). It shows that each stage inherits the conditioning of the one before, up to $O(\sigma)$, so per-stage difficulty need not compound with depth—*provided* the shallow first stage is itself well-conditioned; it is vacuous when that stage is barren. We motivate local strong convexity in Appendix F.5.

3.4 FROM ENERGY TO FIDELITY

Theorem 3.6 (Gap–fidelity sandwich). *If $\Delta = \lambda_2 - \lambda_1 > 0$ and $F(\psi) = |\langle \mathbf{v}_1 | \psi \rangle|^2$, then*

$$\Delta\big(1 - F(\psi)\big) \le \langle \psi | \mathbf{H} | \psi \rangle - \lambda_1 \le (\lambda_d - \lambda_1)\big(1 - F(\psi)\big). \tag{14}$$

Iterating (13) for T steps and splitting the energy error through the depth-$(D{+}2)$ minimizer gives the certificate

$$F(\psi(\boldsymbol{\theta}^{(T)})) \ge 1 - \frac{1}{\Delta}\Big[(1 - \eta\mu_{D+2})^T \big(E_{D+2}(\boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^\star) - E_{D+2}(\boldsymbol{\theta}^\star_{D+2}, \boldsymbol{\phi}^\star)\big) + r_{D+2}\Big], \tag{15}$$

where $r_{D+2} = E_{D+2}(\boldsymbol{\theta}^\star_{D+2}, \boldsymbol{\phi}^\star) - \lambda_1 \ge 0$ is the *reachability residual*. The optimization term vanishes as $T \to \infty$ but r_{D+2} does not: by Theorem 3.2, if $|\mathbf{v}_1\rangle \notin \mathcal{M}_2$ the certificate saturates below 1, and it is likewise uninformative when $\varepsilon \not\ll \Delta$ (near criticality). This is exactly the regime—TFIM below—where the SE ansatz loses.

4 EXPERIMENTS

We organize the experiments as a single line of inquiry. We first verify saturation (Section 4.1), then show that it is what lets added depth help without hurting—no initialization shock (Section 4.2)—isolate the resulting trainability gain (Section 4.3) and price it in gates (Section 4.4), compare against established VQE optimizers across budgets (Section 4.5), and scale to $n = 16$ (Section 4.6). Generality over nine Hamiltonians, the SE-vs-HEA expressibility ceiling, growth schedules, hyperparameter sensitivity, and ablations are reported in Appendix D; we summarize them where relevant.

Setup. Unless stated otherwise: $n = 6$ qubits, exact statevector simulation (PennyLane), mean$\pm$std over 10 seeds. Nine Hamiltonians span a range of entanglement requirements (Table 5, Appendix D.1); the primary three are TFIM ($\lambda_1 = -7.727$, $\Delta = 0.27$; ground state leaves $\mathcal{M}_2$), Tilted Ising (-10.214, 0.44; in $\mathcal{M}_2$), and all-to-all Random Ising (-3.405, 0.53; intermediate). Progressive methods use stage depths $\{1, 3, 5\}$ with T steps per stage; baselines get the same $3T$ steps at fixed depth; $\eta = 0.02$, $\sigma = 0.05$; the shared first block is drawn from the same seed for all methods. "Naïve PDT" is the *copy*-initialized variant (each appended block copies the shared first block), the strongest naïve warm start; the *random*-initialized variant appears only in Section 4.2.

4.1 REACHABLE SET SATURATION

Theorem 3.2 predicts that $\mathcal{M}_D$ stabilizes at $D \ge 2$. We train SE circuits from scratch at $D \in \{1, 2, 3, 5, 7, 11\}$ for 2000 steps (Figure 2). On Tilted Ising the saturation is unambiguous: $D = 1$ stalls at -9.000 ± 0.000, while every $D \ge 2$ reaches -10.200 ± 0.007, within 0.014 of the exact ground state, with the standard deviation collapsing by two orders of magnitude between $D = 1$ and $D = 2$. On TFIM all $D \ge 2$ land on similar means (-7.29 to -7.43) but the ceiling sits well above the exact energy (gap 0.30)—a first glimpse of the expressibility limit for highly entangled ground states. On Random Ising, $D = 5$ posts the best mean (-3.191 ± 0.361) and deeper circuits do not improve on it, again consistent with saturation confounded by local minima. In every case the reachable set has closed by $D = 2$.

4.2 INITIALIZATION SHOCK AND MONOTONICITY

We trace the energy across stages at $T = 200$ for IP-PDT and for naïve PDT with copy and with random initialization (Figure 3). At each expansion (dotted lines) random-initialized naïve PDT spikes as the optimizer scrambles to undo the new CNOT gates—on Tilted Ising it oscillates between -10 and 0 at every expansion—while IP-PDT slides through on a smooth, monotonically non-increasing trajectory to -10.194 ± 0.009, as Theorem 3.4 guarantees. The copy-initialized naïve variant perturbs the trained state far less and nearly coincides with IP-PDT on this trace. What still favors IP-PDT is *structural*: the cancelling pairs collapse the trained circuit to a single entangling layer ($\sim 5\times$ fewer two-qubit gates) and preserve the energy *exactly* at every expansion, whereas copy initialization only makes the shock empirically small with no guarantee.

4.3 TRAINABILITY BEYOND EXPRESSIBILITY

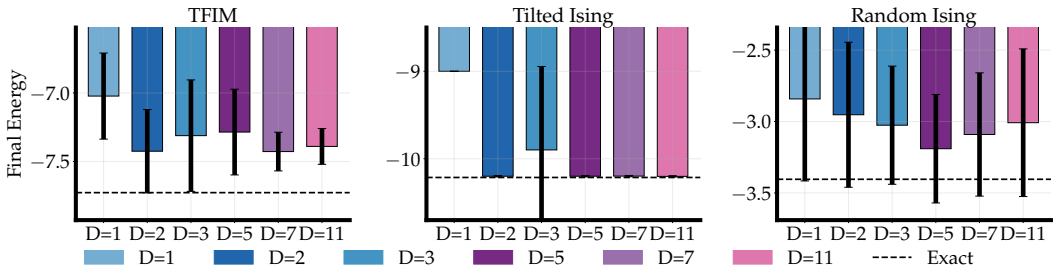


Figure 2: Reachable set saturation: final energy vs. depth for the SE ansatz (2000 steps, 10 seeds). $D = 1$ is strictly worse; $D \geq 2$ saturates, confirming Theorem 3.2. Dashed line: exact ground state.

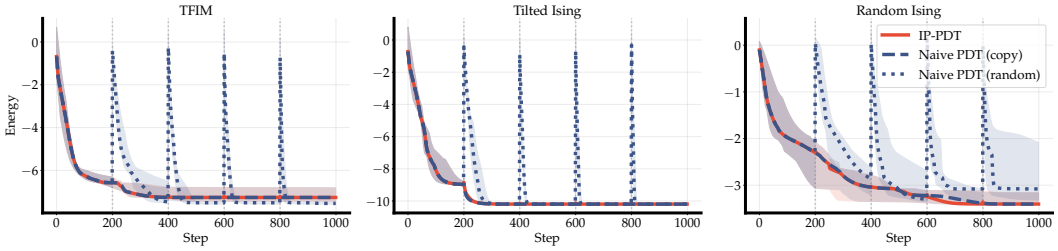


Figure 3: Initialization shock ($T = 200$ per stage, 10 seeds). IP-PDT (red) descends monotonically; random-initialized naïve PDT spikes at each expansion (dotted lines); copy-initialized naïve PDT largely avoids the shock and nearly coincides with IP-PDT, so IP-PDT's advantage over it is the $\sim 5\times$ CNOT saving and the exact guarantee of Theorem 3.4.

Saturation enables the cleanest test of *why* progressive training helps. Since $\mathcal{M}_2 = \mathcal{M}_5$, a depth-5 SE circuit trained from scratch ("SE scratch": all $15n$ parameters at once, one random initialization, same $3T$ budget) can reach exactly the states IP-PDT can, so any gap between them is a *trainability* gap. We pit IP-PDT ($1 \to 3 \to 5$) against SE scratch at $D \in \{1, 2, 5\}$ for $T \in \{100, 200, 500\}$ (Figure 4). On Random Ising IP-PDT matches or beats every scratch depth at $T = 100, 500$ (at $T = 200$ only $D = 5$ edges it) and holds markedly lower variance throughout. On Tilted Ising it matches scratch-$D = 5$ at all budgets (both reach -10.194 by $T = 100$), well ahead of scratch-$D = 1$ (plateau -9.0) and scratch-$D = 2$ (± 0.91 variance). The exception is instructive: on TFIM scratch-$D = 2$ overtakes IP-PDT at $T = 500$ (-7.425 vs. -7.048) because it spends all 1500 steps on its 36

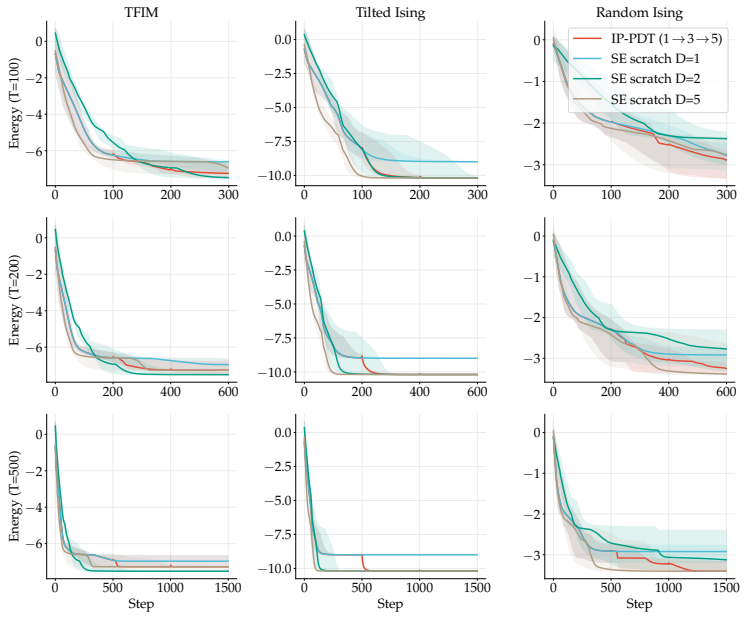


Figure 4: Trainability beyond expressibility. Rows: $T \in \{100, 200, 500\}$; columns: Hamiltonians. All methods share the reachable set $\mathcal{M}_2$. IP-PDT (red) matches or outperforms all scratch depths on Random and Tilted Ising with far lower variance. Median±IQR, 10 seeds.

parameters, whereas IP-PDT spends its first 500 at depth 1, where the expressibility ceiling leaves little to gain.

4.4 Cost-adjusted comparison

Counting steps flatters the SE ansatz, whose single CNOT ring is five times cheaper per layer than HEA's. Plotting energy against cumulative gate count for 300 steps (Figure 5), HEA spends 36,000 gates to IP-PDT's 18,000, yet the cheaper method reaches lower energy on all three primary Hamiltonians: Tilted Ising -10.194 vs. -9.407, Random Ising -2.919 vs. -2.770, TFIM -7.126 vs. -6.948. On the metric that matters for near-term devices, IP-PDT does more with half the gates. Across the full nine-Hamiltonian benchmark at $T = 200$, IP-PDT takes the lowest or tied-lowest energy gap on eight of nine, losing only TFIM to HEA (Table 6); against HEA specifically it wins four of five Hamiltonians ranging from easy to Heisenberg-hard, again ceding only TFIM (Appendix D.3). We conjecture that expressibility and trainability trade against each other: a better-conditioned landscape more than offsets five times fewer entangling gates, except when the target cannot be represented at all.

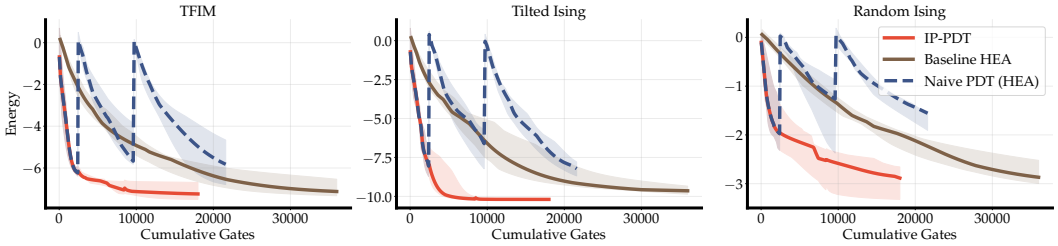


Figure 5: Energy vs. cumulative gate cost (300 steps). IP-PDT (red, 18K gates) reaches lower energy than HEA (brown, 36K gates) on all three Hamiltonians. Median±IQR, 10 seeds.

4.5 Comparison with established VQE methods across budgets

We budget-match IP-PDT against SE scratch ($D = 5$), HEA ($D = 5$), Rotosolve (Ostaszewski et al., 2021), layerwise training (Skolik et al., 2021), and a simplified ADAPT-VQE (Grimsley et al., 2019), sweeping $T \in \{50, 100, 200, 500\}$ (Figure 6; the single-budget view at 300 evaluations is in Appendix D.4). IP-PDT's edge is sharpest where near-term hardware lives—at low budgets. At $T = 50$ it posts the lowest energy on all three Hamiltonians; on Tilted Ising it reaches near-exact energy (-10.19) by $T = 100$ while layerwise and Rotosolve remain above -8.0 at $T = 500$. On Random Ising the advantage holds at every budget. Only on TFIM does a competitor catch up: SE scratch closes the gap at $T = 500$ as both press against the expressibility ceiling. Rotosolve and ADAPT-VQE lag at every budget. Growth schedule, learning rate, and steps-per-stage sweeps (Appendices D.5 and D.6) show a wide, flat optimum, and a jitter ablation over $\sigma \in \{0, 0.01, 0.05, 0.1, 0.5\}$ and zero-initialization (Appendix D.7) shows all progressive variants performing nearly identically and all beating scratch: it is the curriculum, not the jitter, that drives the gain.

4.6 Scaling to $n = 16$

We re-run the nine-Hamiltonian benchmark at $n = 16$ ($d = 65{,}536$; Pauli-sum Hamiltonians, Lanczos ground energies), same protocol (Table 1). *HEA fails to scale*: on every Hamiltonian the depth-5 HEA produces gaps 2–10× larger than the single-entangler methods, consistent with barren-plateau and trap-dominated regimes (McClean et al., 2018; Anschuetz & Kiani, 2022). *Progressive methods remain competitive with full-depth SE* on $6/9$ Hamiltonians despite starting shallow. *IP-PDT and copy-initialized naïve PDT converge at scale*, trading wins of ≤ 0.1–0.3; the jitter centre matters less at $n = 16$. One caveat on error bars: on Tilted Ising, IP-PDT's larger mean and variance are not a single outlier—a 40-seed check finds IP-PDT in a poor basin (gap > 1) on $4/40$ seeds, copy-naïve on $3/40$, with identical medians (0.06), while SE scratch fails on $16/40$; an occasional poor basin is a property of the SE landscape at this scale, shared by both progressive methods and $4\times$ rarer than from scratch. On Spin Glass the large spread is likewise shared by all methods, and IP-PDT's median gap (≈ 0.65) beats naïve PDT (≈ 3.3) and scratch (≈ 4.5).

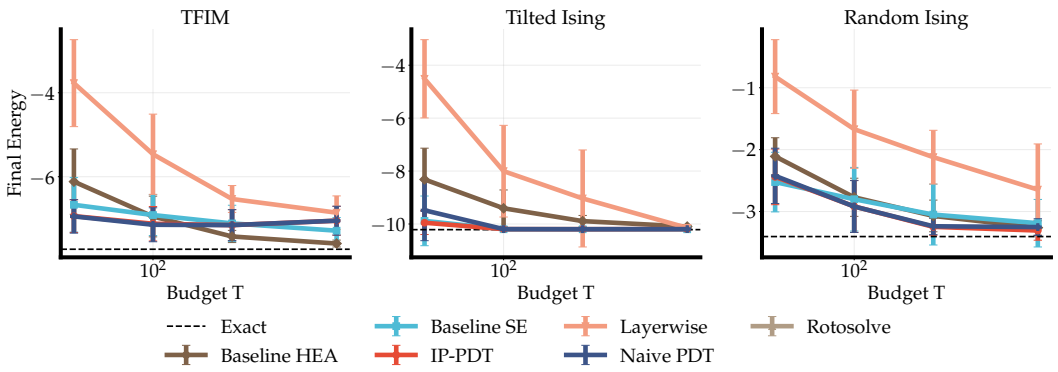


Figure 6: Established VQE methods at multiple budgets. IP-PDT (red) dominates at low budgets; SE scratch catches up at $T = 500$ on TFIM; layerwise and Rotosolve lag at all budgets. Mean±std, 10 seeds. Dashed: exact ground state.

Table 1: Energy gap $E - \lambda_1$ at $n = 16$ (mean±std, 10 seeds). **Bold** = best; † = tied within 1 std.

Hamiltonian	IP-PDT	Naïve PDT	SE scratch	HEA
Tilted Ising	0.356±.955	**0.055**±.017	0.624±1.27	3.684±2.28
Mixed-field	**0.065**±.020†	0.065±.020	0.622±1.28	3.357±2.22
Spin Glass	**3.449**±4.12	4.012±3.89	4.814±2.93	6.003±4.78
Random Ising	2.565±1.76	2.777±1.75	**2.102**±.81	5.861±2.94
Power-law	**0.025**±.001†	0.025±.001	1.500±1.10	3.103±3.41
Random XY	1.917±.69	**1.633**±.48	2.163±1.43	11.44±5.76
XXZ	**1.548**±.82	1.763±.97	1.553±.73	4.184±2.35
Heisenberg	5.140±1.69	4.892±1.49	**4.204**±1.71	9.444±4.72
TFIM	1.654±.66	1.710±.70	**1.218**±.59	1.924±.77

5 Conclusion

IP-PDT grows a variational circuit by forward/reverse block pairs whose CNOT gates cancel, eliminating initialization shock and collapsing the circuit to a single entangling layer. The Saturation Theorem shows this manifold expands once and then freezes, so every later gain is a trainability gain; continuation, monotonicity, and basin-preservation results explain why depth can be spent safely on optimization, and experiments at $n = 6$ and $n = 16$ show IP-PDT matching or beating full-depth HEA with $5\times$ fewer CNOTs, most strongly at low budgets. *Limitations.* (i) IP-PDT is not a barren-plateau escape: for global cost functions all ansätze flatten identically (Cerezo et al., 2021b), and Theorem 3.5 presumes a well-conditioned shallow stage. (ii) The expressibility ceiling binds for strongly correlated targets outside $\mathcal{M}_2$: SE wins when the target lies in $\mathcal{M}_2$ and loses near criticality (TFIM), and on dense-coupling Hamiltonians the full-depth baseline can outperform the progressive strategies, so deeper entangling architectures remain necessary there. (iii) Our guarantees cover descent on previous-stage parameters; joint dynamics are open. Natural extensions are $K > 1$ backbone entangling layers with rotation-only growth, and composing depth continuation with Hamiltonian-space continuation (Žunkovič et al., 2026).

Reproducibility Statement

All theoretical claims are stated with their assumptions in Section 3; complete proofs, including the full continuation and basin-preservation arguments and the finite-shot extension, are given in Appendices C, E and F. Experimental protocol, budget matching, seeds, and hyperparameters are specified in Section 4 and Appendices B.4 and D.1; every Hamiltonian is defined in Table 5. Anonymized source code reproducing all figures and tables is provided as supplementary material.

AI use statement

We used generative AI tools to help write the text of this paper and to help locate papers that needed to be cited. We did not use generative AI tools for any other task, including the theoretical results, proofs, or experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eugene L. Allgower and Kurt Georg. *Introduction to Numerical Continuation Methods*, volume 45 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), 2003.

Eric R. Anschuetz and Bobak T. Kiani. Beyond barren plateaus: Quantum variational algorithms are swamped with traps. *Nature Communications*, 13:7760, 2022.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.

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

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

Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019.

Harper R. Grimsley, Sophia E. Economou, Edwin Barnes, and Nicholas J. Mayhall. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nature Communications*, 10:3007, 2019.

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

Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549:242–246, 2017.

Martín Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3:542–551, 2023.

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

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 10th anniversary edition, 2010.

Mateusz Ostaszewski, Edward Grant, and Marcello Benedetti. Structure optimization for parameterized quantum circuits. *Quantum*, 5:391, 2021.

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

Boris T. Polyak. Gradient methods for minimizing functionals. *Zh. Vychisl. Mat. i Mat. Fiz.*, 3(4): 643–653, 1963. English translation: USSR Comput. Math. Math. Phys. 3(4):864–878, 1963.

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

Rafał Szlendak, Elnur Gasanov, and Peter Richtárik. Understanding progressive training through the framework of randomized coordinate descent. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *PMLR*, pp. 2161–2169, 2024.

Bojan Žunkovič, Marco Ballarin, Lewis Wright, and Michael Lubasch. Scalable, self-verifying variational quantum eigensolver using adiabatic warm starts. arXiv:2602.17612 [quant-ph], 2026.