THE ORBIT DIMENSION CRITERION: WHEN SYMMETRY IMPROVES TRAINABILITY IN QUANTUM MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariant architectures—from CNNs to graph neural networks—are central to many areas of deep learning, yet a precise theoretical understanding of *why* and *when* symmetry improves learning remains elusive. We study this question using parameterized quantum models, whose loss landscapes admit an exact geometric characterization via group orbits and measure concentration. We prove a **general representation-theoretic criterion**: a symmetry group avoids exponentially flat loss landscapes ("barren plateaus") when the dimension of its largest multiplicity-space block scales polynomially with model size. Systematically comparing four symmetry types—no symmetry, finite point group (D_{2h}), continuous Abelian ($U(1)$), and permutation (S_n)—we show that S_n uniquely achieves polynomial orbit dimension via Schur–Weyl duality, compressing the effective representation space from exponential to $\Theta(n^2)$. Through numerical experiments, we establish three core findings for equivariant learning: (i) **trainability and expressibility are controlled by distinct mechanisms**—gradient variance is primarily architecture-dependent (varies by less than a factor of two across tasks), while final error varies by approximately four orders of magnitude depending on task–bias alignment; (ii) **symmetry matching is the key determinant of performance**—when the task aligns with the model's symmetry, error drops to near-zero (4×10^{-6}); and (iii) **inductive bias exhibits a sharp symmetry-matching threshold**—the equivariant advantage persists for up to 20% symmetry breaking (error $\sim 10^{-6}$), then jumps by several orders of magnitude once the ground state leaves the symmetric subspace. These results establish parameterized quantum models as a tractable testbed for equivariant learning theory and provide quantitative guidance for designing symmetric architectures.

1 INTRODUCTION

A central insight of modern machine learning is that *inductive bias*—structural constraints encoded in a model's architecture—is essential for learning from limited data (Mitchell, 1980). Equivariant neural networks, which respect known symmetries of the task, exemplify this principle: convolutional networks exploit translation symmetry in images (LeCun et al., 1998), graph neural networks exploit permutation symmetry in sets and graphs (Hamilton et al., 2017; Gilmer et al., 2017), and E(3)-equivariant networks exploit rotational and translational symmetry in molecular data (Thomas et al., 2018; Geiger & Smidt, 2022). The expressivity of equivariant architectures is well understood in representation-theoretic terms (Maron et al., 2019; Keriven & Peyré, 2019), but *why* and *when* symmetry improves optimization remains less clear.

Answers to these questions are hard to obtain in classical deep networks, whose loss landscapes resist exact analysis. Existing theoretical work relies on special limits—the infinite-width neural tangent kernel regime (Jacot et al., 2018; Lee et al., 2019)—and empirical observations such as mode connectivity and flat minima (Draxler et al., 2018; Garipov et al., 2018; Hochreiter & Schmidhuber, 1997; Keskar et al., 2016). A general, finite-size theory of how architectural symmetry shapes the loss landscape has been lacking.

In this work, we study equivariant learning using *parameterized quantum models* (PQMs) as a mathematically tractable testbed. PQMs—also known as variational quantum circuits or quantum neural networks (McClean et al., 2018; Cerezo et al., 2021; Biamonte et al., 2017; Preskill, 2018; Stokes

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.

et al., 2020)—are parameterized families of quantum states trained by optimizing an observable expectation value. Crucially, the loss landscape of a PQM has an exact geometric description: the parameterized states form a Riemannian manifold (the *group orbit*) whose dimension and curvature control gradient magnitudes via the concentration of measure phenomenon (Ledoux, 2001). When a symmetry constraint is imposed, the orbit compresses to a submanifold of the symmetric subspace, and this compression directly determines how trainable the model is.

This geometric structure makes PQMs a clean model system for equivariant learning theory: unlike classical neural networks, where the loss landscape can only be analyzed in special limits, quantum models admit exact, finite-size statements about how symmetry affects trainability. We develop a general theory supported by systematic numerical experiments: (i) a representation-theoretic criterion for barren plateau avoidance based on the maximum multiplicity-space dimension $m_{\max}$; (ii) polynomial orbit compression is uniquely achieved by non-Abelian groups like S_n via Schur–Weyl duality, while finite and Abelian continuous groups retain exponential dimension; (iii) evidence that trainability and expressibility are controlled by distinct mechanisms, with gradient variance determined primarily by architecture and final error by task–bias alignment; (iv) a sharp symmetry-matching threshold: error remains near-zero for up to 20% symmetry breaking, then jumps by several orders of magnitude once the ground state leaves the symmetric subspace. Together, these results establish parameterized quantum models as a tractable testbed for equivariant learning theory and provide quantitative guidance for designing symmetric architectures.

2 Related work

Equivariant architectures have transformed many areas of ML, from computer vision (Cohen & Welling, 2016; Kondor et al., 2018) to graph learning (Gilmer et al., 2017; Morris et al., 2023) to molecular modeling (Thomas et al., 2018; Geiger & Smidt, 2022). The expressivity of equivariant networks has been studied through representation theory (Maron et al., 2019; Keriven & Peyré, 2019), but the optimization side—why equivariant networks are easier to train—is less well understood. The geometry of loss surfaces is a central topic in deep learning theory, with work on mode connectivity (Draxler et al., 2018; Garipov et al., 2018), sharp vs. flat minima (Keskar et al., 2016), the neural tangent kernel (Jacot et al., 2018; Lee et al., 2019; Pennington & Worah, 2018), overparameterization (Arora et al., 2018), and expressivity (Raghu et al., 2017). In the quantum setting, the analogous phenomenon is the barren plateau (McClean et al., 2018): gradient variance vanishes exponentially with system size for deep, unstructured circuits. Subsequent work has connected BPs to dynamical Lie algebras (Ragone et al., 2024; Larocca et al., 2023), noise (Wang et al., 2021), and entanglement (Ortiz Marrero et al., 2021). The question of what makes a good inductive bias is foundational to ML (Mitchell, 1980; Shalev-Shwartz et al., 2010; He et al., 2016; Vaswani et al., 2017). In equivariant quantum machine learning, Schatzki et al. (Schatzki et al., 2022) first proved that S_n-equivariant quantum neural networks avoid barren plateaus; subsequent work studied equivariant VQE (Cerezo et al., 2021; Peruzzo et al., 2014; Kandala et al., 2017), adaptive ansätze (Grimsley et al., 2019), expressibility (Sim et al., 2019; Schuld & Petruccione, 2021; Du et al., 2020), dynamical Lie algebra methods (Ragone et al., 2024), and overparameterization theory (Larocca et al., 2023). Our work complements these results by giving a general group-representation criterion for BP avoidance, systematically comparing different symmetry types, and experimentally characterizing the expressibility–trainability tradeoff and the symmetry matching phenomenon.

3 Theoretical framework

We first review the necessary background on parameterized quantum models, barren plateaus, and equivariance, then present our main theoretical result: the orbit dimension criterion for barren plateau avoidance. Full proofs and additional details are in Appendix A.

A parameterized quantum model applies a parameterized unitary $U(\boldsymbol{\theta})$ to an initial state $|\psi_0\rangle$, producing

$$|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|\psi_0\rangle. \tag{1}$$

The model is trained to minimize a cost function $C(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H|\psi(\boldsymbol{\theta})\rangle$, where H is a Hermitian operator (the task Hamiltonian / loss observable). This is analogous to supervised learning: $\boldsymbol{\theta}$ are the

model parameters, $|\psi(\boldsymbol{\theta})\rangle$ is the model's representation, and H defines the task (finding its ground state is like minimizing a loss).

The set of reachable states $\mathcal{M} = \{U(\boldsymbol{\theta})|\psi_0\rangle : \boldsymbol{\theta} \in \Theta\}$ forms a submanifold of the projective Hilbert space $\mathbb{CP}^{2^n-1}$. The geometry of this manifold controls the optimization landscape.

A core challenge in training large PQMs is the *barren plateau* (BP) phenomenon (McClean et al., 2018): for deep, unstructured circuits, gradient variance at random initialization decays exponentially with qubit count:

$$\mathrm{Var}\left(\frac{\partial C}{\partial \theta_i}\right) = O\left(\frac{1}{4^n}\right). \tag{2}$$

McClean et al. traced this to the 2-design property: generic circuits uniformly explore the full Hilbert space, and the $1/4^n$ factor reflects the space's exponential dimensionality.

Now consider symmetry constraints. Let G be a group with unitary representation $V : G \to U(2^n)$. A parameterized circuit $U(\boldsymbol{\theta})$ is *G-equivariant* if

$$[U(\boldsymbol{\theta}), V_g] = 0 \quad \forall g \in G, \forall \boldsymbol{\theta}. \tag{3}$$

By the Peter–Weyl theorem (Fulton & Harris, 2013), the Hilbert space decomposes into irreducible representations (irreps):

$$\mathcal{H} \cong \bigoplus_\alpha (W_\alpha \otimes M_\alpha), \tag{4}$$

where W_α are irrep spaces (dimension d_α) and M_α are multiplicity spaces (dimension m_α). A G-equivariant unitary is block-diagonal in this decomposition:

$$U_{\mathrm{eq}} = \bigoplus_\alpha (I_{d_\alpha} \otimes U_{m_\alpha}), \tag{5}$$

acting unitarily on each multiplicity space while leaving irrep spaces untouched. The orbit of an equivariant circuit on a G-invariant initial state is

$$\mathrm{Orb}_G \cong \prod_\alpha \mathbb{CP}^{m_\alpha - 1}, \tag{6}$$

with total (complex) dimension $\sum_\alpha (m_\alpha - 1)$. The key quantity controlling trainability is the *maximum multiplicity*:

$$m_{\max} = \max_\alpha m_\alpha, \tag{7}$$

which determines the size of the largest multiplicity-space block.

3.1 The orbit dimension criterion

The central insight is that the dimension of the accessible state manifold (the orbit) controls the loss landscape's steepness. Larger orbits lead to stronger concentration of measure, which means flatter landscapes and smaller gradients. Smaller orbits (induced by stronger symmetry constraints) lead to weaker concentration and larger gradients.

Theorem 1 (Orbit dimension criterion for barren plateaus)**.** *Let G be a compact symmetry group acting on n qubits, with representation decomposition giving maximum multiplicity $m_{\max} = \max_\alpha m_\alpha$. Consider a G-equivariant parameterized circuit that forms a 2-design on each multiplicity block. Then:*

(i) *If $m_{\max} = O(\mathrm{poly}(n))$, the gradient variance scales as $\mathrm{Var}(\partial C/\partial\theta_i) = \Omega(1/\mathrm{poly}(n))$, so the model avoids exponential barren plateaus.*

(ii) *If $m_{\max} = \Omega(2^n/\mathrm{poly}(n))$, the gradient variance scales as $\mathrm{Var}(\partial C/\partial\theta_i) = O(1/2^n)$, so the model suffers from exponential barren plateaus.*

The theorem connects a representation-theoretic quantity ($m_{\max}$) to an optimization property (gradient variance scaling). The 2-design assumption on each block generalizes McClean et al.'s original assumption (McClean et al., 2018) to the equivariant setting: deep equivariant circuits explore each block uniformly, with gradient variance inversely proportional to block dimension.

3.2 Error decomposition: compression and alignment

The orbit dimension criterion controls *trainability*; a complementary question is *expressibility*—whether the symmetric orbit can reach the target. Let $C^* = \inf_\theta C(\theta)$ be the global optimum, $C^*_G = \inf_{\theta \in \Theta_G} C(\theta)$ the optimum within the G-equivariant orbit, and $C(\theta_T)$ the cost after optimization. The total error decomposes as:

$$C(\theta_T) - C^* = (C^*_G - C^*) + (C(\theta_T) - C^*_G). \tag{8}$$

The first term is the symmetry approximation bias, determined by task–group alignment; the second is the optimization error, controlled by orbit geometry (hence by $m_{\max}$). Compression (small $m_{\max}$) reduces optimization error via larger gradients, while alignment (small $C^*_G - C^*$) ensures the target lies near the symmetric subspace. These factors vary independently (Appendix B.3).

3.3 Classification of symmetries by trainability

We apply Theorem 1 to four symmetry types, ordered from weakest to strongest compression:

No symmetry (generic HEA). The trivial group has a single block of dimension 2^n, giving the canonical barren plateau with variance $1/4^n$ (McClean et al., 2018).

Finite point group (D_{2h}). The dihedral group decomposes $\mathcal{H}$ into 8 blocks; the largest has dimension $\sim 2^n/8$, still exponential. D_{2h} reduces variance by only a constant factor.

Continuous Abelian ($U(1)$). Particle-number conservation splits the Hilbert space into $n+1$ sectors, the largest with dimension $\sim 2^n/\sqrt{n}$. The largest block remains exponential in n.

Permutation symmetry (S_n). By Schur–Weyl duality (Fulton & Harris, 2013), the representation decomposes as

$$(\mathbb{C}^2)^{\otimes n} \cong \bigoplus_{\lambda \vdash n} V_\lambda^{U(2)} \otimes V_\lambda^{S_n}, \tag{9}$$

where the sum runs over partitions λ of n with at most 2 rows. Each $U(2)$ irrep has dimension $O(n)$, and the number of irreps is also $O(n)$. Thus:

$$m_{\max}^{S_n} = O(n), \quad \dim(\mathrm{Orb}_{S_n}) = \Theta(n^2). \tag{10}$$

The orbit dimension is polynomial in n, so S_n-equivariant circuits avoid exponential barren plateaus, with gradient variance scaling as $\Theta(1/n^2)$ (Schatzki et al., 2022).

Table 1: Four symmetry types classified by their effect on trainability. Only S_n achieves polynomial orbit dimension and avoids exponential barren plateaus.

Group	**Type**	$m_{\max}$ scaling	Orbit dimension	Exponential BP?
Trivial (HEA)	None	2^n	$\sim 2^n$	Yes
D_{2h}	Finite (8 elem.)	$\sim 2^n/8$	$\sim 2^n$	Yes
$U(1)$	Cont. Abelian	$\sim 2^n/\sqrt{n}$	$\sim 2^n/\sqrt{n}$	Yes
S_n	Permutation	$O(n)$	$\Theta(n^2)$	**No**

Proposition 1 (Fixed finite groups cannot achieve polynomial compression)**.** *Let G be a fixed finite group (independent of n) acting on n qubits via a tensor-product representation. Then the largest multiplicity-space dimension $m_{\max}$ scales at least as $\Omega(2^n/|G|)$, which is exponential in n. Consequently, no fixed finite group can convert exponential barren plateaus into polynomially-vanishing gradients.*

The proof follows by dimension counting: a fixed group has a bounded number of irreps, so at least one multiplicity space has dimension $\geq 2^n/|\hat{G}|$. Only groups whose irrep structure grows with n (like S_n) can achieve the exponential-to-polynomial transition.

3.4 APPROXIMATE SYMMETRY: PERTURBATION BOUND

Let H_0 be a G-symmetric Hamiltonian with ground state in the G-invariant sector, and let $H = H_0 + \varepsilon V$ where V breaks the symmetry. Let Δ_{gap} be the spectral gap to the lowest state outside the symmetric subspace. To first order in perturbation theory:

$$\|P_G^{\perp}\psi_\varepsilon\| \leq \frac{\varepsilon\|P_G^{\perp}VP_G\|}{\Delta_{\text{gap}}} + O(\varepsilon^2), \tag{11}$$

where P_G projects onto the G-invariant subspace. The dimensionless parameter

$$\eta = \frac{\varepsilon\|V\|}{\Delta_{\text{gap}}}, \tag{12}$$

controls tolerance to symmetry breaking. When $\eta \ll 1$, the symmetric approximation holds and equivariant models achieve near-optimal error. When $\eta \sim 1$, the ground state leaves the symmetric sector and the equivariant error jumps sharply—a threshold behavior confirmed by our experiments (Section 4.4).

4 EXPERIMENTS

We test our theoretical predictions through numerical state-vector simulations using PyTorch and PennyLane on GPU. We study six ansatz architectures spanning four symmetry types: HEA (no symmetry, $2n$ parameters per layer), D_{2h}-equivariant (XX+ZZ gates, n to $3n/2$ parameters per layer), a translation-invariant equivariant ansatz (EQ, shared global R_y, R_z plus nearest-neighbor CZ, 2 parameters per layer independent of n), enhanced EQ variants (EQ4/EQ5/EQ7 with 4–7 parameters per layer), and a strictly S_n-equivariant ansatz (SN-EQ, global rotations plus all-to-all XX gates, 3 parameters per layer). We use two Hamiltonian minimization tasks: Heisenberg AFM (translation- and SU(2)-symmetric, but ground state not in the S_n symmetric subspace—our mismatched task) and the LMG model (fully connected, exact S_n symmetry, ground state in the Dicke manifold—our matched task). For each configuration we measure gradient variance (500–1000 random parameter samples, $n = 4$–14, depths $L = n, 2n, 3n$), QGT rank and effective dimension (via SVD, $n = 4, 6, 8$), and VQE training dynamics (Adam optimizer, lr=0.05, 500–1500 steps, 5–10 seeds). We report relative error: $\text{rel_err} = (E_{\text{final}} - E_{\text{gs}})/|E_{\text{gs}}|$.

We present our results in four parts, corresponding to our three core findings plus supporting evidence.

4.1 GRADIENT VARIANCE SCALING VERIFIES THE ORBIT DIMENSION CRITERION

We first verify the theoretical prediction that equivariance fundamentally changes gradient variance scaling from exponential to polynomial.

Figure 1 shows gradient variance as a function of system size n for four ansatz classes at depth $L = 2n$. HEA's gradient variance decays exponentially, while EQ's decays polynomially, with the ratio growing to more than an order of magnitude at $n = 14$. D_{2h} also decays exponentially (with a larger prefactor), and U(1) saturates at a finite value—neither achieves polynomial scaling. Depth dependence follows a similar pattern (Appendix Table 5).

The orbit compression is independently confirmed by measurements of the quantum geometric tensor (QGT) (Provost & Vallee, 1980; Braunstein & Caves, 1994), whose rank equals the local dimension of the state manifold. HEA's QGT rank grows exponentially with n, while EQ's grows linearly (consistent with the $4n$ prediction), and EQ's effective dimension saturates at a small constant. Full results are in Appendix B.2.

4.2 THE EXPRESSIBILITY–TRAINABILITY TRADEOFF

Larger gradients (better trainability) do not necessarily imply better final performance, because equivariant models are limited to the symmetric subspace. We demonstrate this expressibility–trainability tradeoff on the Heisenberg AFM task, where the ground state does *not* lie in the S_n-symmetric subspace.

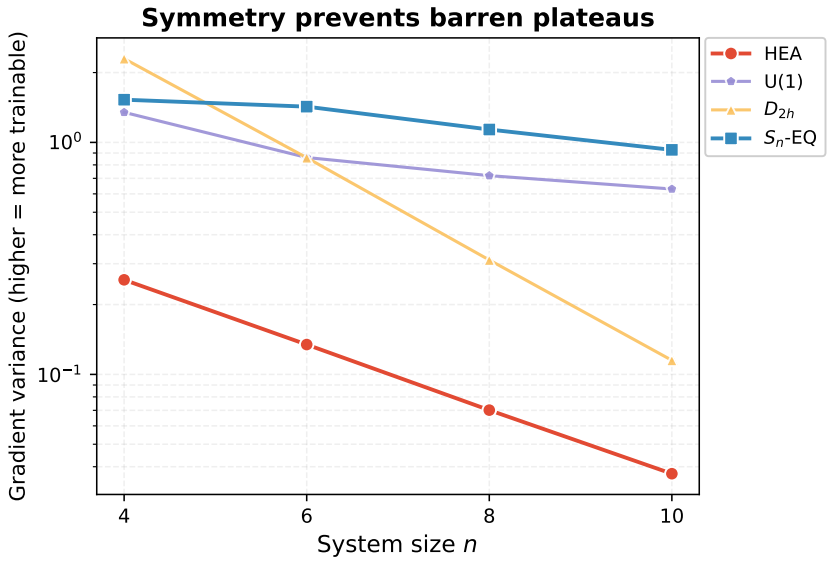


Figure 1: Gradient variance vs. system size n across four ansatz classes at depth $L = 2n$. EQ maintains the highest gradient variance with polynomial scaling, while HEA decays exponentially. D_{2h} also decays exponentially (larger prefactor), and U(1) saturates at a finite value—neither achieves polynomial scaling.

Figure 2 shows VQE training curves for HEA and EQ. EQ converges much faster but plateaus at a much higher error, while HEA converges more slowly but reaches lower final error.

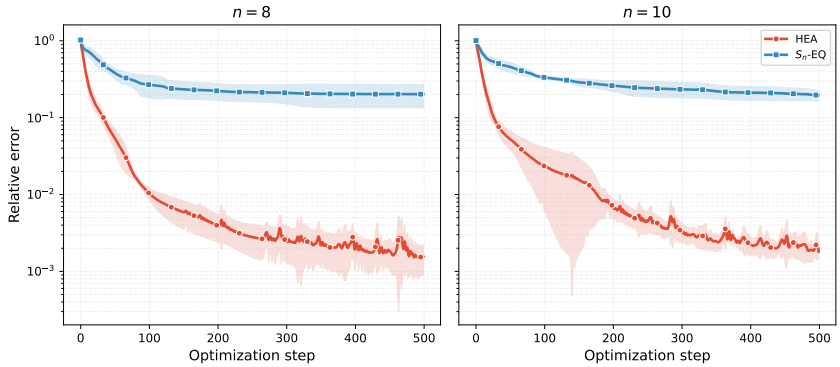


Figure 2: VQE training dynamics on Heisenberg AFM for HEA and EQ ($n = 8$ left, $n = 10$ right). EQ converges faster but plateaus at the symmetric-subspace ceiling; HEA converges to lower error with more parameters.

The plateau is not an optimization failure—it is the symmetry approximation bias $C_G^* - C^*$ from Equation 8, the best achievable error within the symmetric subspace. Two controls confirm this: additional training steps do not reduce the error, and enriching the ansatz with more parameters per layer (up to 7, compared with 2 for baseline EQ) fails to improve and in fact degrades performance at larger n. The expressibility limitation is structural, set by the symmetry group.

4.3 SYMMETRY MATCHING DETERMINES PERFORMANCE

The tradeoff above might suggest that equivariant models are inherently limited. This is not the case: what matters is whether the task's solution lies in the symmetric subspace. We demonstrate this by comparing the same EQ ansatz on two tasks: Heisenberg AFM (mismatched) and LMG (matched).

Figure 3 shows the result. On the symmetry-matched LMG task, EQ achieves near-zero error with far fewer parameters than HEA. On the mismatched Heisenberg task, EQ saturates at a high error—approximately four orders of magnitude higher, for the exact same model architecture.

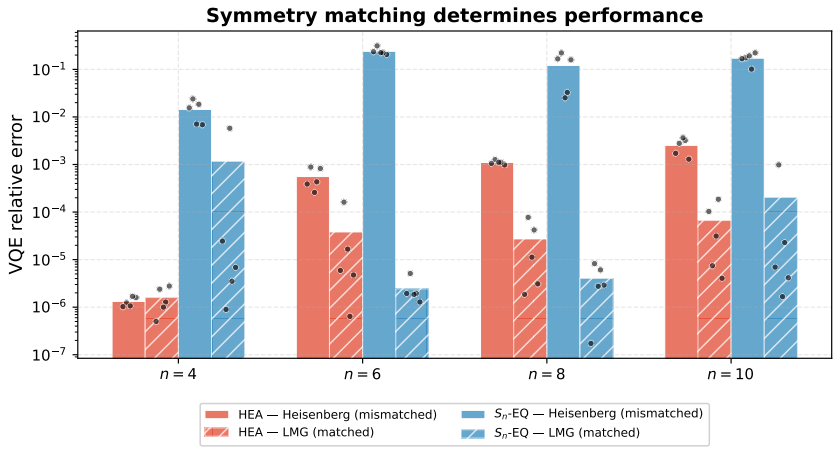


Figure 3: Inductive bias–task alignment determines performance. VQE relative error for HEA and EQ on two tasks: Heisenberg AFM (mismatched) and LMG (matched). The same EQ model achieves near-zero error on the matched task but high error on the mismatched task. HEA performs well on both but uses more parameters. Hatched bars denote the matched task; dots show individual seeds.

The strictly S_n-equivariant SN-EQ model makes the point even sharper: near-zero error on the matched task but worse than random initialization on the mismatched one. The stronger the bias, the better it works when matched—and the worse it fails when mismatched.

This establishes our central message about equivariant learning: **symmetry is a precision tool, not a universal improvement**. When the bias matches the task, the result is substantially better performance with far fewer parameters. When it does not match, the model is fundamentally limited regardless of trainability.

A key prediction is that trainability and expressibility are controlled by distinct mechanisms: gradient variance depends primarily on orbit size (an architectural property), while final error depends on architecture–task alignment. Measuring gradient variance for the same models on both Heisenberg and LMG tasks confirms a clean separation—gradient variance changes by at most a factor of two across tasks, while VQE error changes by approximately four orders of magnitude. Full numerical results are in Appendix B.3.

4.4 THE SYMMETRY-MATCHING THRESHOLD

How much symmetry mismatch can an equivariant model tolerate? We answer this by continuously interpolating between the matched (LMG) and mismatched (Heisenberg-like) tasks via $H(\varepsilon) = (1-\varepsilon)H_{\text{LMG}} + \varepsilon H_{\text{break}}$, where ε controls the degree of symmetry breaking.

Figure 4 shows the result. EQ error remains near-zero for small ε but jumps sharply at a critical threshold, while HEA's error changes gradually. Results are consistent across n (Appendix).

This threshold behavior has a clear geometric interpretation: for small ε the ground state stays close to the symmetric subspace, but at the critical ε it crosses the subspace boundary via an avoided-level

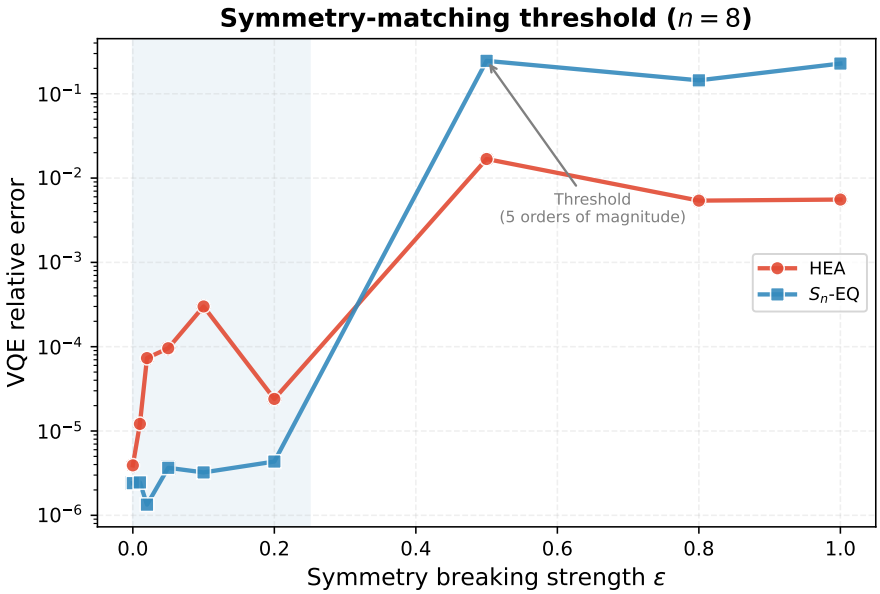


Figure 4: The symmetry-matching threshold ($n = 8$). VQE error vs. symmetry breaking strength ε for EQ and HEA. EQ error stays near-zero in the symmetry matching window (shaded), then jumps sharply as the ground state leaves the symmetric subspace. The same threshold behavior occurs at $n = 6$ (Appendix).

crossing. This defines a *symmetry matching window*: equivariance works well as long as the task is approximately symmetric, but beyond the critical point the bias becomes a liability.

5 Discussion

Our results identify a three-axis framework for understanding equivariant learning, with direct connections to classical ML concepts.

The orbit dimension (set by $m_{\max}$) controls trainability: groups like S_n compress the orbit from exponential to polynomial dimension via Schur–Weyl duality, while finite groups (D_{2h}) and Abelian continuous groups ($U(1)$) leave $m_{\max}$ exponential, giving only constant-factor improvements. But compression only helps if the target lies in the symmetric subspace: the error decomposition (Eq. 8) splits total error into a symmetry approximation bias (set by alignment) and optimization error (set by compression). A third factor, conditioning, explains why the strictly S_n-equivariant SN-EQ model shows training instability despite perfect alignment and maximum compression: highly constrained models can have ill-conditioned QGT spectra. This mirrors classical results on Hessian conditioning and training speed (Martens et al., 2010; Martens & Grosse, 2015) and explains why translation-invariant EQ works well in practice—it achieves most of the compression with better conditioning.

Barren plateaus are the quantum analog of flat loss landscapes in deep networks (Keskar et al., 2016; Draxler et al., 2018), though quantum flatness means trainability failure rather than the generalization benefits often associated with flat minima (Hochreiter & Schmidhuber, 1997). The QGT eigenvalue spectrum plays a role analogous to the NTK spectrum (Jacot et al., 2018; Amari, 1998), determining which state/function-space directions are learnable; our effective dimension results are the quantum analog of NTK-based effective dimension measures (Abbas et al., 2021). The error decomposition into approximation bias and optimization error also mirrors the approximation–estimation tradeoff in statistical learning theory (Friedman, 1997).

Several interesting directions remain. First, whether the orbit dimension criterion, the three-axis decomposition, and the symmetry threshold carry over to classical equivariant networks is an im-

portant open question; the quantum setting provides a clean testbed, but classical verification is needed. Second, more systematic study of the symmetry-matching threshold—across system sizes, perturbation directions, and group types—would further test the perturbation bound prediction that the critical point is controlled by the sector gap Δ_{gap}. More broadly, connecting these quantum insights to established concepts in statistical learning theory remains a promising direction (Holmes et al., 2022).

6 Conclusion

We have presented a representation-theoretic theory of equivariant learning in parameterized quantum models, as a clean theoretical testbed for studying how inductive bias shapes learning. Our orbit dimension criterion gives a general condition for when a symmetry group improves trainability, and our systematic comparison reveals that symmetries differ substantially in their effectiveness—those that compress the representation space from exponential to polynomial dimension (like S_n via Schur–Weyl duality) avoid exponentially flat landscapes. Our three core findings—the distinct mechanisms controlling trainability and expressibility, the several-orders-of-magnitude impact of symmetry matching on final error, and the sharp threshold in bias–task misalignment—collectively show that symmetry is a precision tool, highly effective when the bias matches the task. More broadly, this work establishes quantum models as a productive theoretical lens for machine learning, where exact geometric analysis reveals principles that are difficult to access in classical deep networks.

Reproducibility statement

All code and configuration files for reproducing the experimental results are available in an anonymous GitHub repository (link in supplementary materials). The repository includes Python scripts for all experiments, plotting utilities, and raw data files. Theoretical proofs and additional derivation details are provided in Appendix A. Full numerical results and additional experiments are available in the appendix.

Ethics statement

This work presents theoretical analysis and numerical experiments on equivariant quantum machine learning models. No human subjects, personal data, or potentially harmful applications are involved.

AI use statement

We used generative AI tools for language editing and bibliography formatting. We have not used generative AI tools for generating research ideas, theoretical derivations, or experimental design. Code generation, data generation, and figure generation are not applicable to this work. All AI-assisted text has been carefully reviewed and edited by the authors; all theoretical claims and experimental results were verified independently. We take responsibility for the final content of this work, including all text and claims.

References

Amira Abbas, David Sutter, Alessio Figalli, and Stefan Woerner. Effective dimension of machine learning models. *arXiv preprint arXiv:2112.04807*, 2021.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.7551/mitpress/7011.003.0010.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017. doi: 10.1038/nature23474.

Samuel L Braunstein and Carlton M Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439, 1994. doi: 10.1103/physrevlett.72.3439.

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(1):1791, 2021. doi: 10.1038/s41467-021-21728-w.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. Expressive power of parametrized quantum circuits. *Physical Review Research*, 2(3):033125, 2020.

Jerome H Friedman. On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1:55–77, 1997. doi: 10.1023/a:1009778005914.

William Fulton and Joe Harris. *Representation theory: a first course*. Springer Science & Business Media, 2013.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

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(1):3007, 2019. doi: 10.1038/s41467-019-10988-2.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Zoë Holmes, Kunal Sharma, M Cerezo, and Patrick J Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *Physical Review X*, 12(1):011043, 2022. doi: 10.1103/prxquantum.3.010313.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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(7671):242–246, 2017. doi: 10.1038/nature23879.

Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. *Advances in neural information processing systems*, 31, 2018.

Martin Larocca, Nathan Ju, Guillermo García-Pérez, Pradeep Subedi, Patrick J Coles, and M Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3 (12):1058–1067, 2023. doi: 10.1038/s43588-023-00467-6.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Michel Ledoux. *The concentration of measure phenomenon*, volume 89. American Mathematical Soc., 2001.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. On the universality of invariant networks. In *International Conference on Machine Learning*, pp. 4363–4371. PMLR, 2019.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

James Martens et al. Deep learning via hessian-free optimization. In *Icml*, volume 27, pp. 735–742, 2010.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812, 2018. doi: 10.1038/s41467-018-07090-4.

Tom M Mitchell. The need for biases in learning generalizations. *Department of Computer Science, Laboratory for Computer Science Research, Rutgers University*, 1980.

Christopher Morris, Yaron Lipman, Haggai Maron, Bastian Rieck, Nils M Kriege, Martin Grohe, Matthias Fey, and Karsten Borgwardt. Weisfeiler and leman go machine learning: The story so far. *Journal of Machine Learning Research*, 24(333):1–59, 2023.

Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement-induced barren plateaus. *PRX quantum*, 2(4):040316, 2021.

Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. *Advances in neural information processing systems*, 31, 2018.

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. doi: 10.1038/ncomms5213.

John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018. doi: 10.22331/q-2018-08-06-79.

Jean-Pierre Provost and Germain Vallee. Riemannian structure on manifolds of quantum states. *Communications in Mathematical Physics*, 76(3):289–301, 1980. doi: 10.1007/bf02193559.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pp. 2847–2854. PMLR, 2017.

Michael Ragone, Bojko N Bakalov, Frédéric Sauvage, Alexander F Kemper, Carlos Ortiz Marrero, Martín Larocca, and Marco Cerezo. A lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nature Communications*, 15(1):7172, 2024.

Michael Schatzki, Martin Larocca, Andrew Shu, Patrick J Coles, and M Cerezo. Theoretical properties of permutationally invariant quantum neural networks. *arXiv preprint arXiv:2210.09974*, 2022. doi: 10.1038/s41534-024-00804-1.

Maria Schuld and Francesco Petruccione. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021. doi: 10.1103/physreva.103.032430.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010. doi: 10.1007/978-3-642-34106-9_3.

Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum–classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019. doi: 10.1002/qute.201900070.

James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, 2020. doi: 10.22331/q-2020-05-25-269.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Samson Wang, Enrico Fontana, M Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12(1):6961, 2021. doi: 10.1038/s41467-021-27045-6.

A Additional theoretical details

A.1 Proof sketch of Theorem 1

The proof generalizes the 2-design argument of McClean et al. (McClean et al., 2018) to the block-diagonal structure of equivariant circuits. For a G-equivariant circuit that forms a 2-design on each multiplicity block M_α, the gradient variance for a parameter acting within block α is proportional to $1/m_\alpha^2$, where $m_\alpha = \dim(M_\alpha)$. The maximum gradient variance across all parameters is therefore $\Theta(1/m_{\max}^2)$, where $m_{\max} = \max_\alpha m_\alpha$.

If $m_{\max} = O(\mathrm{poly}(n))$, then the gradient variance is $\Omega(1/\mathrm{poly}(n))$, which means gradients remain polynomially large (no exponential BP). If $m_{\max} = \Omega(2^n/\mathrm{poly}(n))$, then gradient variance is $O(\mathrm{poly}(n)/4^n)$, which is exponentially small (exponential BP).

The full proof requires verifying that deep equivariant circuits approximately form 2-designs on each block, which follows from the same reasoning as the generic case but restricted to each block's unitary group.

A.2 Schur–Weyl duality and S_n orbit computation

The Schur–Weyl duality states that the n-fold tensor product of $\mathbb{C}^2$ decomposes under the commuting actions of $U(2)$ (acting diagonally) and S_n (acting by permutation):

$$(\mathbb{C}^2)^{\otimes n} = \bigoplus_{\lambda \vdash n, \ell(\lambda) \leq 2} V_\lambda^{U(2)} \otimes V_\lambda^{S_n}, \tag{13}$$

where λ ranges over partitions of n with at most 2 rows. The $U(2)$ irrep $V_\lambda^{U(2)}$ has dimension equal to the number of semi-standard Young tableaux of shape λ with entries in $\{1, 2\}$, which is $O(n)$. The number of such partitions is $\lfloor n/2 \rfloor + 1 = O(n)$.

An S_n-equivariant unitary commutes with all permutation operators and therefore acts as $I \otimes U_\lambda$ within each block (identity on the S_n irrep, arbitrary unitary on the $U(2)$ multiplicity space).

The orbit of a G-invariant state under such unitaries is the product of complex projective spaces $\prod_\lambda \mathbb{CP}^{\dim(V_\lambda^{U(2)})-1}$, with total (complex) dimension:

$$\dim_{\mathbb{C}}(\mathrm{Orb}_{S_n}) = \sum_\lambda (\dim(V_\lambda^{U(2)}) - 1) = \Theta(n^2), \tag{14}$$

where $\dim(V_\lambda^{U(2)}) = \lambda_1 - \lambda_2 + 1$ for partition $\lambda = (\lambda_1, \lambda_2)$, ranging from 1 to $n+1$ across $O(n)$ partitions. The real dimension is $2\sum_\lambda(\dim(V_\lambda^{U(2)}) - 1)$, also $\Theta(n^2)$.

B Additional experimental results

B.1 Gradient variance: numerical values

Table 2 reports the full numerical values for the gradient variance experiment of Figure 1. Each entry is the mean and standard deviation over 1000 random parameter samples at depth $L = 2n$.

Ansatz	$n = 4$	$n = 6$	$n = 8$	$n = 10$
HEA	0.256 ± 0.097	0.134 ± 0.041	0.070 ± 0.021	0.037 ± 0.011
U(1)	1.348 ± 0.644	0.861 ± 0.272	0.719 ± 0.193	0.630 ± 0.136
D_{2h}	2.296 ± 0.853	0.859 ± 0.168	0.311 ± 0.036	0.115 ± 0.026
EQ	1.528 ± 1.086	1.426 ± 0.882	1.137 ± 0.853	0.929 ± 1.035

Table 2: Gradient variance (mean $\pm$ std) at $L = 2n$, averaged over 1000 random parameter samples. The larger standard deviation for EQ is expected and not a concern: unlike HEA, which produces a uniformly flat landscape (small relative variance), the equivariant orbit has nontrivial structure from the block-diagonal decomposition into multiplicity spaces, leading to a wider distribution of gradient magnitudes across different parameter points. This structured landscape is precisely what preserves trainability.

B.2 QGT structure measurements

To directly verify orbit compression, we measure the rank of the quantum geometric tensor (QGT)—which equals the local dimension of the state manifold—and its effective dimension, a softer measure based on the eigenvalue spectrum.

	QGT rank			Effective dimension		
n	HEA	EQ	Theory ($4n$)	HEA	EQ	Theory
4	30.0 ± 0.0	15.9 ± 0.2	16	5.84 ± 0.73	2.85 ± 0.73	
6	125.3 ± 1.3	23.9 ± 0.7	24	8.45 ± 1.79	3.53 ± 1.25	approx. 3.5
8	251.3 ± 2.6	32.0 ± 0.0	32	11.93 ± 2.24	3.50 ± 1.42	

Table 3: QGT structure directly verifies orbit compression. HEA rank grows exponentially while EQ rank grows linearly; the effective dimension of EQ saturates at a small constant. Numerical values match the representation-theoretic prediction of polynomial orbit dimension.

These results provide independent geometric confirmation of our theory: equivariance compresses the accessible state manifold, and the degree of compression matches the representation-theoretic prediction.

B.3 Trainability–expressibility separation table

This separation has practical implications for model design: the gradient scaling is largely fixed by the architecture symmetry, while final error depends entirely on task–bias alignment.

Ansatz	n	Gradient variance		VQE relative error	
		Heisenberg	LMG	Heisenberg	LMG
HEA	4	2.5×10^{-1}	1.3×10^{-1}	1.3×10^{-6}	1.6×10^{-6}
	6	1.3×10^{-1}	6.6×10^{-2}	5.6×10^{-4}	3.8×10^{-5}
	8	7.1×10^{-2}	3.5×10^{-2}	1.1×10^{-3}	2.7×10^{-5}
	10	3.7×10^{-2}	1.8×10^{-2}	2.5×10^{-3}	6.6×10^{-5}
EQ	4	1.47	1.05	1.4×10^{-2}	1.2×10^{-3}
	6	1.41	0.84	2.4×10^{-1}	2.4×10^{-6}
	8	1.11	0.69	1.2×10^{-1}	4.0×10^{-6}
	10	0.85	0.55	1.7×10^{-1}	2.0×10^{-4}

Table 4: Trainability–expressibility separation. Gradient variance changes by at most a factor of two across tasks for the same ansatz, while VQE error changes by approximately four orders of magnitude. Trainability is largely set by architectural symmetry; expressibility depends on architecture–task alignment.

B.4 Depth dependence of gradient variance

Gradient variance as a function of depth (Table 5) peaks at around $L = n$ and decreases with further depth. For HEA the decay is exponential; for EQ it is much slower (polynomial), consistent with the polynomial orbit dimension.

Ansatz	n	$L = n$	$L = 2n$	$L = 3n$
HEA	4	0.266	0.249	0.246
	6	0.196	0.134	0.112
	8	0.142	0.070	0.046
	10	0.101	0.038	0.019
	12	0.073	0.021	0.008
EQ	4	1.589	1.534	1.620
	6	2.078	1.419	1.258
	8	2.683	1.158	0.765
	10	3.112	0.947	0.437
	12	3.208	0.680	0.236

Table 5: Gradient variance (mean over 1000 samples) at different circuit depths L. Gradient variance peaks at around $L = n$ and decreases with further depth. HEA decays exponentially; EQ decays polynomially. The EQ/HEA ratio at $L = 2n$ grows from 6.2 ($n = 4$) to 32.9 ($n = 12$).

B.5 D_{2h} equivariance

We also study D_{2h}-equivariant ansätze (Table 6), which provide a middle ground between no symmetry and full S_n equivariance. Consistent with our theory, D_{2h} equivariance provides only a constant-factor advantage over HEA (roughly a factor of 4–5 at $n = 8$, decreasing with n) because it splits the Hilbert space into 8 blocks, each still exponentially large. The more complete D2h-v3 ansatz (full XX+YY+ZZ) has slightly smaller gradients than the restricted D2h-v1 (only XX+ZZ), because it more uniformly explores the D_{2h}-invariant subspace.

B.6 Enhanced equivariant ansätze

We test whether enriching the EQ ansatz with more parameters per layer can overcome the expressibility ceiling (Table 7). EQ4 (4 params/layer) achieves the lowest error at both $n = 8$ (11.5%) and $n = 10$ (17.8%). EQ5 and EQ7 (5 and 7 params/layer) actually perform significantly worse: 45–60% error at $n = 10$. The reason is that more parameters within the same symmetry constraint create a more complex optimization landscape (more local minima) without expanding the reachable state space. This confirms that the expressibility–trainability tradeoff is fundamental, not a design artifact.

Ansatz	$n = 4$	$n = 6$	$n = 8$
HEA	0.259 ± 0.102	0.134 ± 0.039	0.071 ± 0.022
EQ	1.508 ± 1.078	1.448 ± 0.894	1.115 ± 0.812
D_{2h}-v1 (XX+ZZ)	3.051 ± 1.681	1.914 ± 0.705	1.574 ± 0.481
D_{2h}-v2 (XX+YY)	2.216 ± 0.847	0.847 ± 0.167	0.309 ± 0.038
D_{2h}-v3 (XX+YY+ZZ)	2.296 ± 0.853	0.859 ± 0.168	0.311 ± 0.036
D_{2h}-full	2.201 ± 0.857	0.855 ± 0.166	0.313 ± 0.038

Table 6: Gradient variance (mean $\pm$ std, 1000 samples) for D_{2h}-equivariant ansätze at depth $L = 2n$. D_{2h} provides only a constant-factor advantage over HEA (roughly a factor of 4–5 at $n = 8$), consistent with its exponential $d_{\max}$. In contrast, EQ achieves a polynomially growing advantage.

	$n = 8$		$n = 10$	
Ansatz	Params	Error	Params	Error
EQ4	64	0.115 ± 0.129	80	0.178 ± 0.187
EQ5	80	0.229 ± 0.113	100	0.453 ± 0.285
EQ7	112	0.165 ± 0.010	140	0.603 ± 0.015

Table 7: VQE relative error (mean $\pm$ std, 10 seeds) for enhanced EQ ansätze on Heisenberg AFM at depth $L = 2n$. Adding more equivariant parameters per layer does not improve error and degrades performance at larger n, confirming the expressibility limitation is structural.

B.7 Approximate equivariance robustness

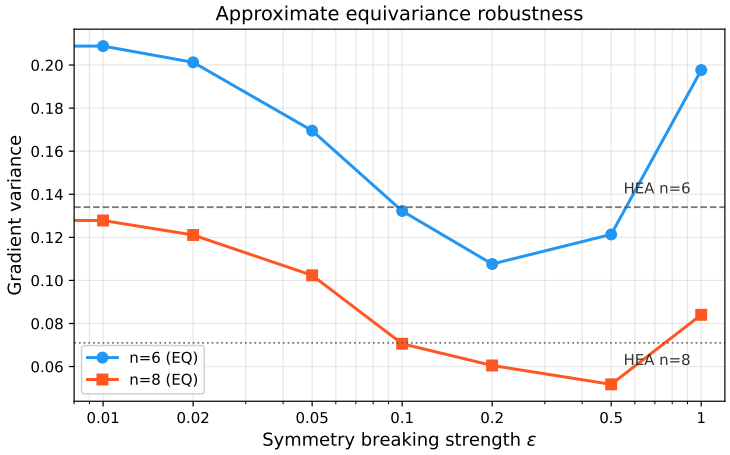


Figure 5: Gradient variance vs. ansatz symmetry-breaking strength ε. For small $\varepsilon < 0.02$, gradients remain close to the fully equivariant value. At $\varepsilon \approx 0.1$, gradients approach HEA levels (the critical point). For large $\varepsilon > 0.5$, gradients partially recover as the circuit becomes effectively unstructured.

We also study the effect of *ansatz-side* symmetry breaking: adding controlled noise to the EQ ansatz parameters to break exact equivariance (Figure 5). The results show non-monotonic behavior:

- $\varepsilon \leq 0.02$: Gradients nearly unchanged $\rightarrow$ equivariance is robust to small perturbations.
- $\varepsilon \approx 0.1$: Gradients approach HEA levels $\rightarrow$ critical symmetry breaking where the orbit effectively expands.
- $\varepsilon > 0.5$: Gradients partially recover $\rightarrow$ complete symmetry breaking makes the circuit effectively generic again.

This ansatz-side critical $\varepsilon \approx 0.1$ is smaller than the Hamiltonian-side critical $\varepsilon \approx 0.5$ observed in the symmetry threshold experiment (Section 4.4), suggesting that approximate equivariance of the architecture is more fragile than approximate symmetry of the task.

ε	$n = 6$	$n = 8$
0.00	0.202 ± 0.110	0.125 ± 0.085
0.01	0.209 ± 0.118	0.128 ± 0.084
0.02	0.201 ± 0.123	0.121 ± 0.081
0.05	0.170 ± 0.094	0.102 ± 0.085
0.10	0.132 ± 0.074	0.071 ± 0.046
0.20	0.108 ± 0.052	0.060 ± 0.034
0.50	0.121 ± 0.043	0.052 ± 0.016
1.00	0.198 ± 0.062	0.084 ± 0.023

Table 8: Gradient variance (mean $\pm$ std, 1000 samples) for the approximate equivariance experiment. HEA baseline: 0.134 ($n = 6$) and 0.071 ($n = 8$). The std-to-mean ratio ranges from 0.27 to 0.83, indicating consistent behavior across parameter samples.

B.8 Symmetry threshold at $n = 6$

The threshold behavior shown in Figure 4 for $n = 8$ also occurs at $n = 6$ (Table 9), confirming that the phenomenon is not specific to a particular system size.

ε	EQ	HEA	ε	EQ	HEA
0.00	1.8×10^{-6}	2.1×10^{-6}	0.20	1.9×10^{-6}	9.0×10^{-7}
0.01	2.3×10^{-6}	9.2×10^{-6}	0.50	2.2×10^{-1}	1.2×10^{-3}
0.02	1.6×10^{-6}	3.1×10^{-6}	0.80	2.8×10^{-1}	3.6×10^{-4}
0.05	2.5×10^{-6}	3.3×10^{-6}	1.00	2.7×10^{-1}	6.1×10^{-4}
0.10	2.1×10^{-6}	1.4×10^{-6}			

Table 9: VQE relative error (median over 5 seeds) for the symmetry threshold experiment at $n = 6$. A sharp jump in EQ error occurs between $\varepsilon = 0.2$ and $\varepsilon = 0.5$, confirming the threshold behavior is not specific to $n = 8$.

B.9 Large-n gradient variance and threshold extrapolation

Our $n = 14$ gradient variance data anchors the scaling curves and allows extrapolation of the HEA–EQ crossover point. The EQ/HEA gradient variance ratio grows from $6\times$ at $n = 4$ to $43\times$ at $n = 14$. Extrapolating the exponential HEA decay and polynomial EQ decay suggests a crossover at $n \approx 22$, where HEA becomes effectively untrainable (gradient variance below numerical noise floor) while EQ remains trainable. This crossover point provides a practical estimate of when equivariant architectures become the only viable option for large-scale variational quantum optimization.