

Trainability and Tangent-Space Concentration in Parameterized Quantum Circuits

Gabriel Mejía¹, Eileen Kuehn¹, Melvin Strobl¹, and Achim Streit¹

Karlsruhe Institute of Technology, Scientific Computing Center, Karlsruhe, Germany
gabriel.ruiz@kit.edu eileen.kuehn@kit.edu melvin.strobl@kit.edu achim.streit@kit.edu

1 Introduction

Quantum machine learning (QML) [4] has emerged as one of the most promising candidates in the NISQ era [16], with the potential to provide quantum advantages [7,1]. However, despite significant effort from the scientific community, finding QML models with practical quantum advantage over classical counterparts has remained elusive. One of the main reasons is the relation between trainability and dequantization [8,6]: provably trainable models may sometimes admit efficient classical approximations [20,21,18,11], while highly expressive models may suffer from barren plateaus, where gradients vanish exponentially with system size [15,13,10,3].

In this work, we analyze trainability from the perspective of observable orbits induced by parameterized quantum circuits (PQCs). We study how the number of independent data-visible tangent directions affects the conditioning of the quantum Fisher information matrix (QFIM) and the quantum neural tangent kernel (QNTK). Our main hypothesis is that trainability is determined not only by the total dimension of the dynamical Lie algebra, but also by how the ansatz generators sample tangent directions at finite depth and by how these directions are observed by the chosen observable, input state, and training data. Thus, two ansätze may generate the same dynamical Lie algebra while producing different finite-depth tangent spectra and, consequently, different QFIM or QNTK condition numbers.

Under this framework, we also analyze how increasing the number of layers either increases or concentrates this visibility by measuring the *novelty* contributed by each layer. Our results suggest that trainability is not controlled only by the dimension of the reachable Lie algebra, but by the spectral distribution of observable-visible tangent directions. In particular, circuits with richer ansatz generators can improve the condition number by adding independent directions that strengthen weak eigenmodes rather than concentrating tangent energy only in dominant directions.

2 Methodology and Results

To characterize the trainable directions visible to the loss function, we use the observable orbit of a PQC, its tangent directions, and a novelty measure that quantifies how much each layer contributes beyond previous layers.

2.1 Observable Orbits and Tangent Novelty

Let $\mathcal{A} = (A_1, \dots, A_R)$ be the Hermitian generators of a PQC. The dynamical Lie algebra generated by the circuit is

$$\mathfrak{g} = \text{Lie}_{\mathbb{R}}(iA_1, \dots, iA_R). \quad (1)$$

with corresponding Lie group $G = \exp(\mathfrak{g})$. For a data-dependent circuit $U(x, \theta)$, the observable orbit is

$$\mathcal{O}_O^G = \{V^\dagger O V : V \in G\}, \quad O_x(\theta) = U^\dagger(x, \theta) O U(x, \theta). \quad (2)$$

For each parameter θ_μ , define the effective generator $K_\mu(x, \theta) = iU^\dagger(x, \theta)\partial_{\theta_\mu}U(x, \theta)$. The observable tangent direction is

$$\frac{\partial O_x(\theta)}{\partial \theta_\mu} = i[K_\mu(x, \theta), O_x(\theta)]. \quad (3)$$

Thus, a parameter is visible to the observable only if this commutator is nonzero. If $[K_\mu, O_x] = 0$ for all x and θ , then the corresponding observable tangent direction and loss-gradient component vanish. For the model output $f_\theta(x) = \text{Tr}[\rho O_x(\theta)]$, we define the data-visible tangent matrix

$$\Phi_{s\mu} = \frac{\partial f_\theta(x_s)}{\partial \theta_\mu} = i \text{Tr}[\rho[K_\mu(x_s, \theta), O_{x_s}(\theta)]]. \quad (4)$$

The QNTK on the training set is $K_{\text{QNTK}} = \Phi\Phi^\top$. For the squared loss $\mathcal{L}(\theta) = \frac{1}{2} \sum_s r_s^2$, with $r_s = f_\theta(x_s) - y_s$, the gradient satisfies

$$\nabla_\theta \mathcal{L} = \Phi^\top \mathbf{r}, \quad \|\nabla_\theta \mathcal{L}\|^2 = \mathbf{r}^\top K_{\text{QNTK}} \mathbf{r}. \quad (5)$$

Therefore, trainability depends not only on the number of visible directions, but also on the QNTK spectrum and its alignment with the residual [14,17,22].

To make this explicit, let $K_{\text{QNTK}} = \sum_{a=1}^r \lambda_a u_a u_a^\top$, with $\lambda_a > 0$. Then

$$\|\nabla_\theta \mathcal{L}\|^2 = \sum_{a=1}^r \lambda_a (u_a^\top \mathbf{r})^2. \quad (6)$$

Large gradients arise when the residual has support along large-eigenvalue tangent directions, while residual components aligned with weak eigenmodes are harder to learn. Equivalently,

$$\mathbf{r}^\top K_{\text{QNTK}} \mathbf{r} = \|K_{\text{QNTK}}\|_F \|\mathbf{r}\mathbf{r}^\top\|_F \mathcal{A}(K_{\text{QNTK}}, \mathbf{r}\mathbf{r}^\top), \quad (7)$$

where

$$\mathcal{A}(K_{\text{QNTK}}, \mathbf{r}\mathbf{r}^\top) = \frac{\text{Tr}[K_{\text{QNTK}} \mathbf{r}\mathbf{r}^\top]}{\|K_{\text{QNTK}}\|_F \|\mathbf{r}\mathbf{r}^\top\|_F}. \quad (8)$$

Thus, the gradient second moment is controlled by both the QNTK spectral scale and its alignment with the task residual. When $\mathbb{E} \nabla_{\boldsymbol{\theta}} \mathcal{L} = 0$ over random initializations, the total gradient variance reduces to

$$\text{Var}[\nabla_{\boldsymbol{\theta}} \mathcal{L}] = \mathbb{E}[\mathbf{r}^\top K_{\text{QNTK}} \mathbf{r}]. \quad (9)$$

Hence, vanishing gradients correspond to a small residual-weighted QNTK second moment, which may result either from *small tangent eigenvalues* or from *poor alignment between the residual and the observable-visible tangent subspace*.

If the nonzero eigenvalues of K_{QNTK} are $\lambda_1 \geq \dots \geq \lambda_r > 0$, then $r = \text{rank}(\Phi)$ counts the number of independent visible directions and $\kappa(K_{\text{QNTK}}) = \lambda_1/\lambda_r$ measures their spectral balance. We also use the effective rank dimension

$$d_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}. \quad (10)$$

In addition, we compute the QFIM condition number to compare observable-visible output geometry with the underlying state-space geometry [19,1,12]. The QFIM captures state-space sensitivity before projection onto the observable and dataset, whereas the QNTK captures the part of this sensitivity visible through the measured output.

To quantify whether additional layers add new directions on the observable orbit, we compute an observable-level novelty based on Hilbert–Schmidt tangent vectors. For an input x_a and parameter θ_μ , let

$$A_\mu(x_a) = \text{vec}(\partial_{\theta_\mu} O_{x_a}(\boldsymbol{\theta})).$$

For layer ℓ , $T_O^{(\ell)}$ denotes the matrix formed by stacking the vectors $A_\mu(x_a)$ for all parameters θ_μ in layer ℓ and all input x_a . Let $P_{\ell-1}^{(O)}$ be the orthogonal projector onto the span of all previous observable tangents, $\text{span}\{T_O^{(1)}, \dots, T_O^{(\ell-1)}\}$. We define

$$\mathcal{N}_\ell^{(O)} = \frac{\|(I - P_{\ell-1}^{(O)})T_O^{(\ell)}\|_F^2}{\|T_O^{(\ell)}\|_F^2}. \quad (11)$$

Thus, $\mathcal{N}_\ell^{(O)}$ measures the fraction of new observable-tangent energy contributed by layer ℓ , using standard projection and subspace-angle geometry [5,9,2].

2.2 Numerical results

We consider a supervised scalar regression task on a one-dimensional input domain. For each input x_s , the PQC predicts the expectation value $f_{\boldsymbol{\theta}}(x_s) = \text{Tr}[\rho U^\dagger(x_s, \boldsymbol{\theta}) O U(x_s, \boldsymbol{\theta})]$, and the loss is the mean-squared error $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{s=1}^N (f_{\boldsymbol{\theta}}(x_s) - y_s)^2$. The labels y_s are generated from a fixed reference circuit and then standardized to zero mean and unit variance.

The observable is a normalized random Pauli Hamiltonian $O = \sum_{m=1}^{12} c_m P_m$, with $\sum_{m=1}^{12} c_m^2 = 1$. We use $N_{\text{train}} = 32$ training points, $N_{\text{test}} = 64$ test points, and depths $L = 1, \dots, 10$. Each ansatz is evaluated with and without data re-uploading.

We compare four ansatz families, denoted HEA, CRX, Diverse, and Symmetric. The HEA ansatz uses local generators $\{Y_i, Z_i\}_{i=1}^n$ together with a fixed entangling structure. The CRX ansatz uses local generators $\{Y_i, Z_i\}_{i=1}^n$ and trainable controlled-rotation generators with Pauli support $\{X_j, Z_i X_j\}_{1 \leq i < j \leq n}$. The Diverse ansatz uses the generator set $\{X_i, Y_i, Z_i\}_{i=1}^n \cup \{X_i X_j, Y_i Y_j, Z_i Z_j\}_{1 \leq i < j \leq n}$. The Symmetric ansatz uses translation-tied generators $\{\sum_i X_i, \sum_i Y_i, \sum_i Z_i, \sum_i X_i X_{i+1}, \sum_i Y_i Y_{i+1}, \sum_i Z_i Z_{i+1}\}$, so each parameter controls an entire translated orbit of local or nearest-neighbor terms.

For each ansatz and depth, we compute the data-visible tangent matrix Φ , the empirical QNTK $K_{\text{QNTK}} = \Phi \Phi^\top$, the QFIM, the QNTK rank r , the QNTK condition number $\kappa(K_{\text{QNTK}})$, the QFIM condition number $\kappa(F_Q)$, the effective rank dimension d_{eff} , the spectral concentration ratio k_{90}/r , and the layer-wise novelty.

Table 1 summarizes the final-layer results at $L = 10$ with $n = 5$ qubits. All ansätze reach the same QNTK rank $r = 32$, the same Lie-algebra dimension $\dim(\mathfrak{g}) = 1023$, and the same observable-orbit dimension $\dim(\mathcal{O}_O) = 960$. Nevertheless, their QNTK condition numbers vary from 5.16 to 581.95, showing that the Lie-algebra dimension alone does not determine the finite-depth trainability geometry. Figure 1 shows the observable novelty $\mathcal{N}_\ell^{(O)}$ across layers. Its ansatz-dependent decay indicates that increasing depth does not uniformly add new observable-orbit directions; rather, each circuit template reuses or expands the visible tangent space at a different rate.

The best-conditioned case is the data-reuploading **diverse_xyz_all** ansatz, with $\kappa(K_{\text{QNTK}}) = 5.1605$, $d_{\text{eff}} = 26.7247$, corresponding to $d_{\text{eff}}/r = 83.51\%$. For this ansatz, $k_{90} = 26$, so $k_{90}/r = 26/32 = 81.25\%$. Therefore, the best-conditioned case is the least spectrally concentrated one. Instead, its tangent energy is distributed across a large fraction of the available observable-visible directions. This suggests that improved trainability is associated with balanced tangent spectra, where weak eigenmodes are strengthened rather than all tangent energy concentrating in a few dominant directions.

Ansatz	Reup.	p_ℓ	$\kappa(K)$	$\kappa(F_Q)$	d_{eff}/r	k_{90}/r	$\bar{\mathcal{N}}^{(O)}$
Diverse	Yes	45	5.16	8.74	83.51%	81.25%	0.8507
CRX	Yes	20	11.15	50.61	69.62%	75.00%	0.8071
HEA	Yes	10	16.19	93.90	69.04%	71.88%	0.9628
Symmetric	Yes	6	145.43	2.22×10^7	52.34%	59.38%	0.8559
Diverse	No	45	166.13	8.68	46.57%	53.13%	0.6476
CRX	No	20	245.03	40.63	41.57%	50.00%	0.6811
HEA	No	10	268.49	66.26	44.53%	50.00%	0.9149
Symmetric	No	6	581.95	2.68×10^5	36.15%	43.75%	0.7898

Table 1: Final-layer diagnostics at $L = 10$ for $n = 5$ qubits. All rows have QNTK rank $r = 32$, Lie-algebra dimension $\dim(\mathfrak{g}) = 1023$, and observable-orbit dimension $\dim(\mathcal{O}_O) = 960$. Here p_ℓ is the number of trainable parameters per layer, $\kappa(K)$ is the QNTK condition number, $\kappa(F_Q)$ is the mean QFIM condition number, d_{eff}/r is the participation-rank percentage, k_{90}/r is the fraction of eigenvalues needed to capture 90% of the QNTK spectral energy (less percentage means more concentration on few eigenmodes), and $\bar{\mathcal{N}}^{(O)}$ is the mean observable novelty across layers.

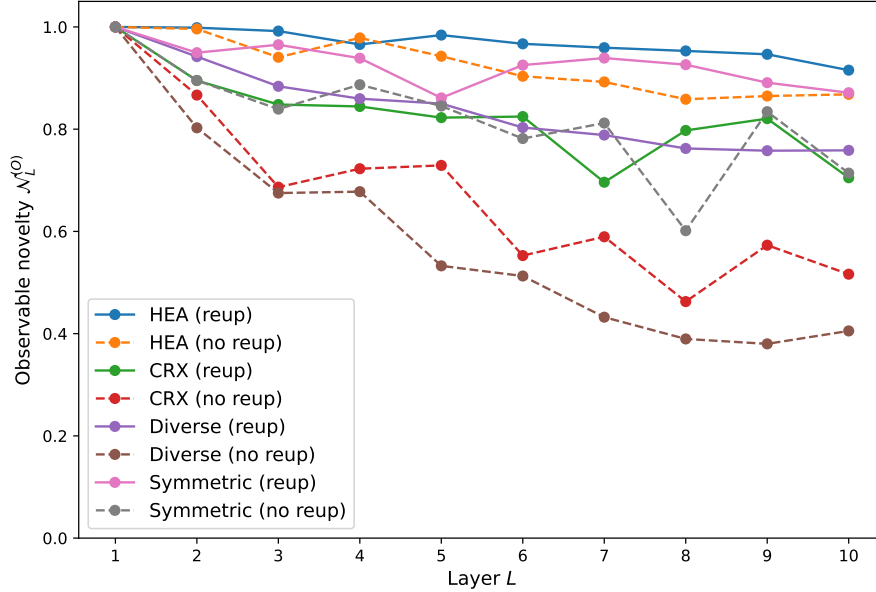


Fig. 1: Observable novelty $\mathcal{N}_L^{(O)}$ as a function of depth for the tested ansätze. Solid lines denote runs with data re-uploading and dashed lines runs without data re-uploading. Higher values indicate that the newly added layer contributes directions that remain more visible at the observable level.

3 Conclusion

We analyzed trainability through the observable-orbit tangent geometry of PQCs. The main conclusion is that the degrees of freedom relevant for training are not determined only by the dimension of the dynamical Lie algebra. What matters is how the ansatz-generated tangent directions become visible through the chosen observable, input state, and data. This visibility is captured by the data-visible tangent matrix Φ , whose rank measures the number of independent visible directions and whose spectrum determines the conditioning of the QNTK.

Our results show that circuits with the same dynamical Lie algebra can have different observable-visible tangent spectra and condition numbers. This occurs because the Lie algebra determines the directions reachable in principle, while the ansatz generator determines how those directions are sampled at finite depth and how strongly they are seen by the observable. Thus, two circuits may have the same reachable algebra but induce different effective training geometries after projection onto the measured output.

The best-conditioned case is the data-reuploading Diverse ansatz. This indicates that improved trainability is associated with a broad and balanced distribution of tangent energy over many observable-visible directions. By contrast, more concentrated spectra may require fewer eigenvalues to capture most of the energy, but they can also lead to worse conditioning because many directions remain weak or poorly visible to the observable.

The observable-novelty analysis shows that increasing depth does not uniformly add new directions on the observable orbit. Later layers may contribute genuinely new observable tangents, but they may also reuse directions already generated by previous layers. Therefore, depth can lead to tangent-space concentration rather than useful expansion of the trainable space. Overall, these results suggest that avoiding barren plateaus is not only a question of reducing expressibility or restricting the Lie algebra. Useful trainability requires an ansatz that generates enough independent observable-visible directions, distributes tangent energy across them without excessive concentration, and aligns those directions with the task-dependent residual.

References

1. Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., Woerner, S.: The power of quantum neural networks. *Nature Computational Science* **1**(6), 403–409 (2021). <https://doi.org/10.1038/s43588-021-00084-1>, <https://doi.org/10.1038/s43588-021-00084-1>
2. Absil, P.A., Mahony, R., Sepulchre, R.: *Optimization Algorithms on Matrix Manifolds*. Princeton University Press (2008), <https://press.princeton.edu/books/hardcover/9780691132983/optimization-algorithms-on-matrix-manifolds>
3. Arrasmith, A., Holmes, Z., Cerezo, M., Coles, P.J.: Equivalence of quantum barren plateaus to cost concentration and narrow gorges. *Quantum Science and Technology* **7**(4), 045015 (2022). <https://doi.org/10.1088/2058-9565/ac7d06>, <https://doi.org/10.1088/2058-9565/ac7d06>
4. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S.: Quantum machine learning. *Nature* **549**(7671), 195–202 (2017). <https://doi.org/10.1038/nature23474>, <https://doi.org/10.1038/nature23474>
5. Björck, Å., Golub, G.H.: Numerical methods for computing angles between linear subspaces. *Mathematics of Computation* **27**(123), 579–594 (1973). <https://doi.org/10.1090/S0025-5718-1973-0348991-3>, <https://doi.org/10.1090/S0025-5718-1973-0348991-3>
6. Cerezo, M., Larocca, M., García-Martín, D., Diaz, N.L., Braccia, P., Fontana, E., Rudolph, M.S., Bermejo, P., Ijaz, A., Thanasilp, S., Anschuetz, E.R., Holmes, Z.: Does provable absence of barren plateaus imply classical simulability? *Nature Communications* **16**, 7907 (2025). <https://doi.org/10.1038/s41467-025-63099-6>, <https://doi.org/10.1038/s41467-025-63099-6>
7. Cerezo, M., Verdon, G., Huang, H.Y., Cincio, L., Coles, P.J.: Challenges and opportunities in quantum machine learning. *arXiv preprint arXiv:2303.09491* (2023). <https://doi.org/10.48550/arXiv.2303.09491>, <https://arxiv.org/abs/2303.09491>
8. Gil-Fuster, E., Gyurik, C., Pérez-Salinas, A., Dunjko, V.: On the relation between trainability and dequantization of variational quantum learning models. In: *Proceedings of the 13th International Conference on Learning Representations* (2025), <https://openreview.net/forum?id=3c215225324f9988858602dc92219615>
9. Golub, G.H., Van Loan, C.F.: *Matrix Computations*. Johns Hopkins University Press, Baltimore, 4 edn. (2013), <https://www.press.jhu.edu/books/title/10678/matrix-computations>
10. Holmes, Z., Sharma, K., Cerezo, M., Coles, P.J.: Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum* **3**, 010313 (2022). <https://doi.org/10.1103/PRXQuantum.3.010313>, <https://doi.org/10.1103/PRXQuantum.3.010313>
11. Landman, J., Thabet, S., Dalyac, C., Mhiri, H., Kashefi, E.: Classically approximating variational quantum machine learning with random fourier features. *arXiv preprint arXiv:2210.13200* (2022). <https://doi.org/10.48550/arXiv.2210.13200>, <https://arxiv.org/abs/2210.13200>
12. Larocca, M., Ju, N., García-Martín, D., Coles, P.J., Cerezo, M.: Theory of overparametrization in quantum neural networks. *Nature Computational Science* **3**(6), 542–551 (2023). <https://doi.org/10.1038/s43588-023-00467-6>, <https://doi.org/10.1038/s43588-023-00467-6>
13. Larocca, M., Thanasilp, S., Wang, S., Sharma, K., Biamonte, J., Coles, P.J., Cincio, L., McClean, J.R., Holmes, Z., Cerezo, M.: Barren plateaus in variational quantum computing. *Nature Reviews Physics* **7**, 174–189 (2025). <https://doi.org/10.1038/s42254-025-00841-7>, <https://doi.org/10.1038/s42254-025-00841-7>
14. Liu, J., Tacchino, F., Glick, J.R., Jiang, L., Mezzacapo, A.: Representation learning via quantum neural tangent kernels. *PRX Quantum* **3**, 030323 (2022). <https://doi.org/10.1103/PRXQuantum.3.030323>, <https://doi.org/10.1103/PRXQuantum.3.030323>
15. McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. *Nature Communications* **9**, 4812 (2018). <https://doi.org/10.1038/s41467-018-07090-4>, <https://doi.org/10.1038/s41467-018-07090-4>
16. Preskill, J.: Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018). <https://doi.org/10.22331/q-2018-08-06-79>, <https://doi.org/10.22331/q-2018-08-06-79>
17. Scala, F., Zoufal, C., Gerace, D., Tacchino, F.: Towards practical quantum neural network diagnostics with neural tangent kernels. *npj Quantum Information* (2026). <https://doi.org/10.1038/s41534-026-01280-5>, <https://doi.org/10.1038/s41534-026-01280-5>, article in press
18. Schreiber, F.J., Eisert, J., Meyer, J.J.: Classical surrogates for quantum learning models. *Physical Review Letters* **131**, 100803 (2023). <https://doi.org/10.1103/PhysRevLett.131.100803>, <https://doi.org/10.1103/PhysRevLett.131.100803>
19. Stokes, J., Izaac, J., Killoran, N., Carleo, G.: Quantum natural gradient. *Quantum* **4**, 269 (2020). <https://doi.org/10.22331/q-2020-05-25-269>, <https://doi.org/10.22331/q-2020-05-25-269>
20. Sweke, R., Recio-Armengol, E., Jerbi, S., Gil-Fuster, E., Fuller, B., Eisert, J., Meyer, J.J.: Potential and limitations of random Fourier features for dequantizing quantum machine learning. *Quantum* **9**, 1640 (2025). <https://doi.org/10.22331/q-2025-02-20-1640>, <https://doi.org/10.22331/q-2025-02-20-1640>
21. Sweke, R., Shin, S., Gil-Fuster, E.: Kernel-based dequantization of variational QML without random fourier features. *arXiv preprint arXiv:2503.23931* (2025). <https://doi.org/10.48550/arXiv.2503.23931>, <https://arxiv.org/abs/2503.23931>
22. You, X., Chakrabarti, S., Chen, B., Wu, X.: Analyzing convergence in quantum neural networks: Deviations from neural tangent kernels (2023), <https://arxiv.org/abs/2303.14844>