Grid-Based Initialization Resolves Frequency Reachability in Trainable-Frequency Quantum Machine Learning

Michael Poppel *michael.poppel@aqarios.com*
LMU Munich & Aqarios GmbH, Munich, Germany

Markus Baumann *markus.baumann@ifi.lmu.de*
LMU Munich, Munich, Germany

Sebastian Wölckert *sebastian.woelckert@ifi.lmu.de*
LMU Munich, Munich, Germany

Claudia Linnhoff-Popien *linnhoff@ifi.lmu.de*
LMU Munich, Munich, Germany

Jonas Stein *jonas.stein@ifi.lmu.de*
LMU Munich, Munich, Germany

arXiv:2602.23409v2 [cs.LG] 15 May 2026

Abstract

Angle-encoded variational quantum circuits admit a truncated Fourier series representation of their output, but approximating functions with maximum frequency $\omega_{\max}$ using fixed unary encoding requires $\mathcal{O}(\omega_{\max})$ encoding gates. Trainable-frequency (TF) circuits promise a reduction by learning the data-encoding prefactors alongside the ansatz parameters, adapting the accessible frequency spectrum to the target during training. We identify a practical barrier that prevents this promise from being realized: the prefactor gradient is suppressed by the spectral gap between the circuit's accessible frequencies and the target spectrum, independently of the ansatz parameters, confining gradient-driven prefactor movement to a narrow neighborhood of initialization. We propose *ternary grid initialization*—setting prefactors to $\{1, 3, 9, \ldots, 3^{k-1}\}$—which resolves this limitation by ensuring every target frequency within $[-\omega_{\max}, \omega_{\max}]$ lies within $\frac{1}{2}$ unit of a grid point at initialization, removing the spectral gap suppression by construction. On a synthetic benchmark with target frequencies shifted well beyond the standard initialization range, ternary initialization achieves median $R^2 = 0.997$ versus 0.18 for unary initialization, with 100% of runs achieving $R^2 > 0.95$ against 0%. CMA-ES with 20× the evaluation budget reaches only 25% success, confirming the limitation is a property of the optimization landscape rather than of gradient-based optimization specifically. Real-world validation on two benchmark datasets demonstrates consistent advantages over both fixed and trainable unary baselines.

1 Introduction

A key theoretical foundation for quantum machine learning is the Fourier representation of angle-encoded variational quantum circuits (VQCs) (Schuld et al., 2021): any such circuit's output admits a truncated Fourier series representation $f_\theta(x) = \sum_{\omega \in \Omega} c_\omega(\theta)\, e^{i\omega x}$, where the accessible frequency set Ω is determined by the circuit's encoding architecture and the Fourier coefficients $c_\omega(\theta)$ are controlled by the trainable ansatz parameters θ. This connects quantum models to harmonic analysis and makes their expressivity precise: approximating a target function requires the accessible spectrum Ω to cover the target's spectral support, and Pérez-Salinas et al. (2021) and Yu et al. (2022) establish that angle-encoded circuits can achieve this universally. Crucially, the spectrum Ω arises directly from the quantum circuit structure, placing Fourier-based modeling at the center of the case for QML as a natural setting for spectral methods (Belis et al., 2026).

Yu et al. (2022) established that incorporating trainable parameters into the classical data pre-processing stage of PQCs — including the encoding prefactors α that determine the accessible frequency spectrum — enables universal approximation with significantly reduced encoding gate count, whereas fixed unary encoding requires a circuit whose size grows with the frequency range needed to approximate the target to a given precision. Building on this, Jaderberg et al. (2024) proposed an explicit trainable-frequency (TF) architecture in which prefactors are learned alongside the ansatz parameters θ via gradient descent, adding spectrum optimization as an additional degree of freedom alongside coefficient learning. Rather than pre-specifying a frequency grid, a TF circuit adapts its accessible spectrum $\Omega(\alpha)$ to the target during training, promising a reduction in encoding gate count from $\mathcal{O}(\omega_{\max})$ to $\mathcal{O}(k_{\text{opt}})$ for unary TF circuits, where k_{opt} is the number of distinct target frequencies, independent of their magnitudes.

We identify and formally characterize a practical barrier that prevents this promise from being realized in general: *frequency reachability*. The gradient $\partial\mathcal{L}/\partial\alpha_j$ is driven by the inner product between the target residual and a probe function whose spectral support is confined to the circuit's current accessible spectrum $\Omega(\alpha)$. When $\Omega(\alpha)$ is spectrally distant from the target support Ω_h, this inner product is suppressed by the reciprocal of the spectral gap $\delta(\alpha, \Omega_h)$ — a structural consequence of the circuit's Fourier architecture that we formalize in Proposition 1. While Jaderberg et al. (2024) demonstrate successful frequency learning on several target functions, we show that reliable convergence requires the target spectrum to lie within the locally reachable neighborhood of initialization — a condition that warrants explicit analysis and motivates the structured initialization strategy we develop here.

To resolve the reachability limitation, we propose *ternary grid initialization*: setting prefactors to $\{1, 3, 9, \ldots, 3^{k-1}\}$, which by the known dense coverage property of ternary encodings (Shin et al., 2023; Peters & Schuld, 2023) places every target frequency within $[-\omega_{\max}, \omega_{\max}]$ within $\frac{1}{2}$ unit of a grid point at initialization. The required prefactor movement therefore never exceeds $\frac{1}{2}$ unit by construction, and the spectral gap suppression identified in Proposition 1 is formally removed since $\delta(\alpha^{(0)}, \Omega_h) \leq \frac{1}{2}$, placing reliable convergence within reach for every seed and target without requiring prior knowledge of the target spectrum.

Contributions.

1. We prove that the target-driven component of the prefactor gradient in TF quantum circuits is suppressed by the spectral gap $\delta(\alpha, \Omega_h)$ between the accessible and target spectra, independently of the ansatz parameters θ (Proposition 1, Section 4.2).
2. We show empirically that reachability failure persists under derivative-free optimization: CMA-ES achieves only 25% success at 20× the evaluation budget, and random initialization spanning the target range produces a strongly bimodal outcome (Section 6).
3. We propose *ternary grid initialization*, prove it formally removes the spectral gap suppression (Corollary 2), and establish its $\mathcal{O}(\log_3 \omega_{\max})$ encoding gate complexity (Section 5).
4. We validate the approach on synthetic and real-world benchmarks, demonstrating consistent advantages over fixed and trainable unary baselines across all settings tested (Section 6).

2 Related Work

Fourier analysis of variational quantum circuits. The Fourier series interpretation of angle-encoded VQCs was established by Schuld et al. (2021), who showed that the accessible frequency set Ω is determined entirely by the eigenvalues of the data-encoding generators, and that the Fourier coefficients are controlled by the ansatz parameters θ. Pérez-Salinas et al. (2021) demonstrated that a single qubit with repeated data re-uploading can universally approximate continuous functions, and Pérez-Salinas et al. (2025) subsequently characterized the minimal circuit architectures achieving this, providing the gate complexity results that we extend to the trainable-frequency setting in Section 5. Holzer & Turkalj (2024) analyzed spectral invariance and maximality properties of QNN frequency spectra, and Belis et al. (2026) argue more broadly that spectral methods — those that learn or manipulate the Fourier spectrum of a model — arise naturally from quantum computation, motivating the precise understanding of frequency coverage developed here.

Spectral training dynamics. The relationship between a circuit's frequency spectrum and its training dynamics has received increasing attention. Duffy & Jastrzebski (2026) proved that spectral bias in PQCs — the tendency to learn low-frequency components faster than high-frequency ones — is governed by the *redundancy* of each frequency: the number of distinct terms in the circuit's Fourier decomposition contributing to that frequency component. Frequencies with higher redundancy receive stronger gradient signals and converge faster, a result that holds both as an upper bound and in expectation under small-angle initialization. The expected gradient results build on the PQC formalism of Wiedmann et al. (2024), which provides a systematic decomposition of circuit outputs into variational polynomial terms. Our analysis of prefactor gradients in Section 4.2 applies related Fourier-analytic tools to the encoding parameters α rather than the ansatz parameters θ, where the gradient structure takes a distinct form arising from the explicit data-dependence of the encoding gates.

Exponential and ternary encodings. Shin et al. (2023) introduced exponential data encoding for quantum supervised learning, establishing that prefactors $\{1, 3, 9, \ldots\}$ generate a dense integer frequency spectrum covering all integers up to $(3^k - 1)/2$ with k encoding gates. Peters & Schuld (2023) showed that ternary encodings achieve the maximum frequency cardinality among separable single-qubit encoding generators. These results establish the spectral coverage properties that underpin the ternary grid initialization proposed here; our contribution is identifying that the same properties resolve the reachability limitation in the trainable-frequency setting.

Trainable-frequency models. Yu et al. (2022) established that incorporating trainable parameters into the classical data pre-processing stage of PQCs, including the encoding prefactors, enables universal approximation with significantly reduced encoding gate count, though this flexibility introduces ambiguity about whether expressivity gains originate from the classical or quantum components of the model. Jaderberg et al. (2024) proposed an explicit trainable-frequency architecture in which prefactors are learned via gradient descent alongside the ansatz parameters, adding spectrum optimization as an additional degree of freedom alongside coefficient learning. We analyze the conditions under which this optimization succeeds, showing that reliable convergence requires the target spectrum to lie within the locally reachable neighborhood of initialization, and propose a structured initialization strategy that ensures this condition holds by construction.

Classical frequency learning. The problem of adapting a model's frequency content to a target is not unique to quantum computing. Random Fourier Features (Rahimi & Recht, 2007) approximate shift-invariant kernels by projecting inputs onto randomly drawn frequencies, with the distribution of frequencies determined by the kernel's Fourier transform. Learned variants (Yang et al., 2015) optimize the frequency distribution directly, playing a role analogous to trainable prefactors in TF circuits. The reachability problem studied here has no direct classical analogue in this setting: classical frequency learning operates on continuous distributions over frequencies rather than a small number of discrete prefactors, and gradient suppression from spectral gap structure does not arise in the same form. The quantum setting's distinctive constraint is precisely the small number of encoding gates — and hence prefactors — that TF circuits must work with, making initialization geometry critical in a way that does not apply to over-parameterized classical models.

Trainability, barren plateaus, and initialization. The trainability of VQCs has been extensively studied through the lens of barren plateaus, where gradient variances decay exponentially with circuit width (McClean et al., 2018). Larocca et al. (2024) provide a comprehensive Lie-algebraic treatment of this phenomenon. Frequency reachability is a distinct trainability limitation: it affects the encoding prefactors specifically and does not reduce to variance scaling of the ansatz gradients, remaining present even in circuits where barren plateaus are absent. Zhang et al. (2025) proposed adaptive Gaussian initialization for ansatz parameters to mitigate barren plateaus; the frequency reachability problem and its resolution via ternary initialization are orthogonal to ansatz initialization and the two strategies are fully compatible.

3 Background

We briefly establish notation and recall the results on which the paper builds. Throughout, $x \in \mathbb{R}$ denotes a scalar input, $\theta \in \mathbb{R}^p$ the ansatz parameters, $\alpha \in \mathbb{R}^k$ the encoding prefactors for univariate inputs, and M a

Hermitian observable with operator norm $\|M\|_{\text{op}} = \sup_{\|v\|=1} \|Mv\|$. For multivariate inputs $x \in \mathbb{R}^d$, prefactors extend naturally to $\alpha \in \mathbb{R}^{k\times d}$; the theoretical results in this paper are stated for the univariate case, and the multivariate experiments in Section 6.5 apply the same initialization strategy per feature. We model the target as a truncated Fourier series $h(x) = \sum_{\omega\in\Omega_h} h_\omega\, e^{i\omega x}$ with finite spectral support $\Omega_h \subset \mathbb{R}$ and coefficients $h_\omega \in \mathbb{C}$; real-world targets are treated as approximately satisfying this assumption within the frequency range of the model. Training minimizes the mean-squared loss $\mathcal{L}(\theta,\alpha) = \frac{1}{2\pi}\int_{-\pi}^{\pi}(f(x,\theta,\alpha) - h(x))^2\,\mathrm{d}x$, proportional to the squared $L^2[-\pi,\pi]$ norm of the residual. We write $K(\alpha) := |\Omega(\alpha)|$ for the number of accessible frequencies at prefactor values α.

3.1 Variational Quantum Circuits as Fourier Series

Definition 1 (Unary Fixed-Frequency VQC)**.** *An L-layer unary fixed-frequency VQC with angle encoding consists of feature maps $S(x) = e^{ixH}$ where $H = \frac{1}{2}\sigma$ for a Pauli operator σ, implementing rotation gates $R(x) = e^{-i\frac{x}{2}\sigma}$ with unit prefactors, parametrized ansatz layers $W_i(\theta_i)$, and an observable M. The circuit interleaves feature maps and ansatz layers:*

$$U(x,\theta) = W_L(\theta_L)\,S(x)\cdots S(x)\,W_0(\theta_0), \tag{1}$$

with output $f_\theta(x) = \langle 0|U^\dagger(x,\theta)\,M\,U(x,\theta)|0\rangle$.

Theorem 1 (Fourier Representation, Schuld et al. 2021)**.** *For a VQC with angle encoding, the output $f_\theta(x)$ is a truncated Fourier series:*

$$f_\theta(x) = \sum_{\omega\in\Omega} c_\omega(\theta)\, e^{i\omega x}, \tag{2}$$

where Ω is determined by the feature map structure and $|\Omega| = 2L+1$ for L-layer unary circuits.

Definition 2 (Ternary Fixed-Frequency VQC)**.** *A ternary fixed-frequency VQC uses exponentially-spaced prefactors: the i-th feature map takes the form $S_i(x) = e^{i\cdot 3^i xH}$ for $i \in \{0,\ldots,L-1\}$. For k encoding gates, this generates frequency spectrum $\Omega_k = \{n \in \mathbb{Z} : |n| \le (3^k-1)/2\}$, covering all integers in $[-(3^k-1)/2, (3^k-1)/2]$ (Shin et al., 2023; Peters & Schuld, 2023).*

Definition 3 (Trainable-Frequency VQC)**.** *A trainable-frequency (TF) VQC extends the fixed-frequency architecture by parameterizing the feature map prefactors: the j-th encoding gate takes the form $S_j(x,\alpha_j) = e^{i\alpha_j x H_j}$, where $\alpha_j \in \mathbb{R}$ is trainable alongside the ansatz parameters θ. The accessible spectrum $\Omega(\alpha)$ therefore depends on the current prefactor values and evolves during training.*

3.2 Universal Approximation

We call ansatz layers *universal* if they are capable of realizing any unitary in $\mathrm{SU}(2^n)$, providing sufficient expressivity to independently control all accessible Fourier coefficients (Schuld et al., 2021).

Theorem 2 (Universal Approximation for TF VQCs, Yu et al. 2022; Pérez-Salinas et al. 2021)**.** *For a unary TF circuit with universal ansatz layers, for any continuous $f : [0,1]^d \to [-1,1]$ and $\varepsilon > 0$, there exist depth L and parameters θ,α achieving $|f_{\theta,\alpha}(x) - f(x)| < \varepsilon$ for all x.*

The encoding gate complexity of ternary TF circuits relative to fixed unary encoding is established formally in Corollary 3.

3.3 Spectral Redundancy and Training Dynamics

Duffy & Jastrzebski (2026) proved that gradient signals for ansatz parameters in PQCs are governed by the *redundancy* $R(\omega)$ of each frequency component, defined as the number of distinct variational terms in the circuit's Fourier decomposition contributing to that frequency. For any parameter θ and frequency ω, the gradient of the loss satisfies

$$|\partial_\theta \mathcal{L}(\omega)| \le 4R(\omega)\|M\|_{\text{tr}}\,|c_\omega^\Delta|, \tag{3}$$

where $\|M\|_{\text{tr}}$ denotes the trace norm, $c_\omega^\Delta = c_\omega(\theta) - h_\omega$ is the coefficient residual at frequency ω, and the bounds in Section 4.2 are stated in terms of $\|M\|_{\text{op}}$ following a different derivation path. Frequencies with

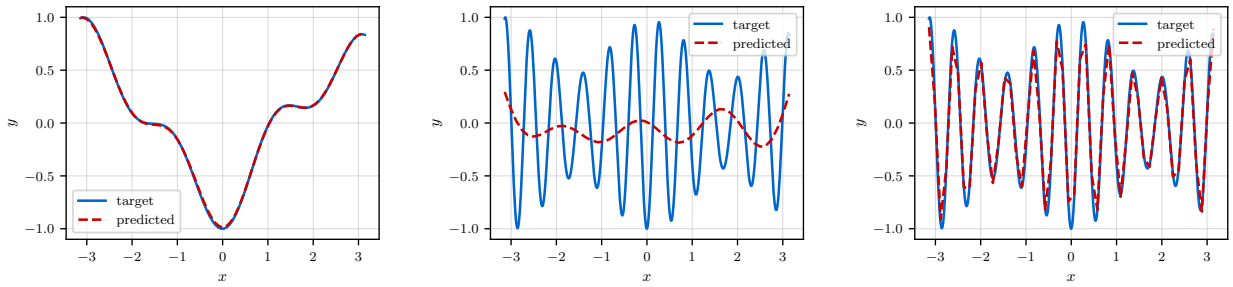
higher redundancy can therefore induce larger gradients and converge faster. Under unary encoding, $R(\omega)$ decays with frequency magnitude, suppressing gradient signals for high-frequency components. Holzer & Turkalj (2024) showed that ternary encoding achieves near-uniform redundancy at the initial grid values $\alpha^{(0)} = \{1, 3, 9, \ldots\}$, with $R(\omega) \in \{1, 2\}$ throughout the accessible spectrum, avoiding the decay that penalizes high frequencies under unary encoding. Once prefactors deviate from consecutive powers of three during training, this uniformity is no longer guaranteed; it is a property of the initialization rather than of the trained circuit. Our analysis in Section 4.2 applies related Fourier-analytic tools to the encoding prefactors α rather than the ansatz parameters θ, where the gradient takes a distinct form arising from the explicit data-dependence of the encoding gates and is characterized by the commutator identity of Lemma 1.

4 Frequency Reachability in Trainable-Frequency Models

The theoretical appeal of TF circuits rests on the assumption that gradient-based optimization can drive frequency prefactors α to arbitrary target values during training. We examine this assumption systematically, first through experiment and then through formal analysis.

4.1 Empirical Characterization

A stress test of long-range frequency trainability. Jaderberg et al. (2024) demonstrate successful fitting of a target function with frequency spectrum $\Omega_1 = \{1, 1.2, 3\}$ using three unary trainable prefactors initialized to 1.0. On inspection, the target frequencies lie within 0.2 units of the initialization — a displacement consistent with the locally reachable neighborhood characterized in Corollary 1 — so this experiment does not test the circuit's ability to learn frequencies far from initialization. To isolate this capability, we shift the target spectrum by 10 units: $\Omega_2 = \{11, 11.2, 13\}$. The circuit architecture, optimizer, and all hyperparameters follow the same general approach as Jaderberg et al. (2024); only the target frequency range changes. As shown in Figure 1(b), optimization fails completely under this shift.



(a) $\Omega_1 = \{1, 1.2, 3\}$ with unary prefactors.

(b) $\Omega_2 = \{11, 11.2, 13\}$ with unary prefactors.

(c) $\Omega_2 = \{11, 11.2, 13\}$ with ternary initialization.

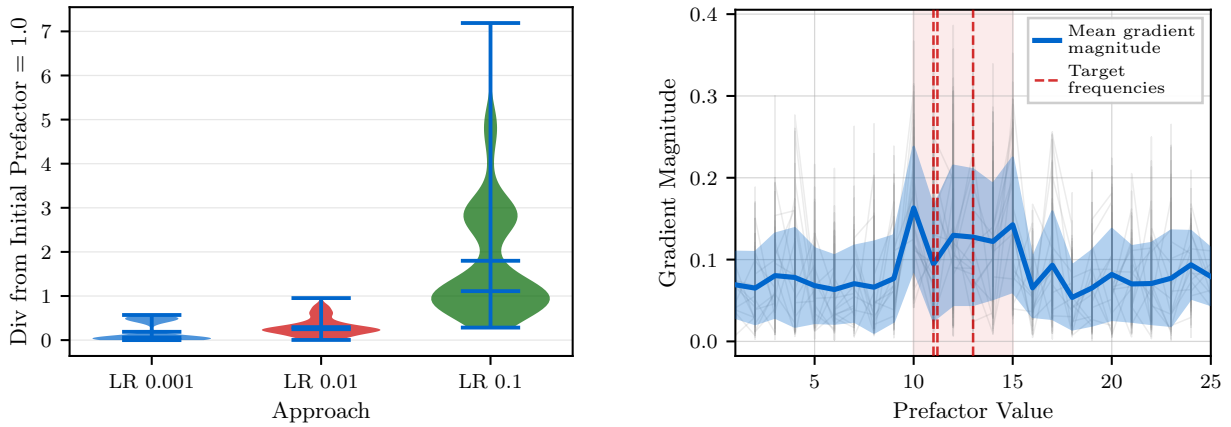
Figure 1: Trained predictions against target for the original Jaderberg et al. (2024) frequency spectrum with unary prefactors (a), the frequency-shifted variant with unary prefactors (b), and the same shifted spectrum with ternary grid initialization (c). Ternary initialization recovers accurate fits where unary fails entirely.

Setup. We use 3-qubit circuits with 3 feature maps and SpecialUnitary ansatz blocks (63 parameters each), following the general approach of Jaderberg et al. (2024). Ten target functions are generated per frequency configuration (Fourier coefficients $a_i, b_i, c_0 \sim \text{Uniform}([0, 1))$), each trained with 10 random weight initializations (100 runs per configuration). Prefactors are initialized to $\alpha^{(0)} = \{1.01, 1.02, 1.03\}$ to ensure distinct gradient signals from the first step and avoid lock-step evolution; Appendix A confirms that the failure mode is insensitive to this choice, with identical initialization $\{1.0, 1.0, 1.0\}$ producing the same failure distribution. All experiments use Adam with learning rate $\eta = 0.001$ for 5,000 steps.

Prefactor displacement. Figure 2(a) quantifies prefactor mobility across learning rates. At standard rates $\eta \in \{0.001, 0.01\}$, both mean and maximum displacement remain below 1 unit after 5,000 steps — far

short of the ≈ 10 units required to reach Ω_2. Even at the aggressive rate $\eta = 0.1$, maximum displacement reaches only ≈ 7 units, with most prefactors still failing to converge. The individual prefactor trajectories underlying these distributions are shown in Appendix A (Figure 9).

Gradient locality. Figure 2(b) reveals the underlying cause. Prefactor gradient magnitudes $|\partial\mathcal{L}/\partial\alpha_i|$, measured at initialization across a sweep of initial prefactor values, are elevated near the target frequency range $\Omega_2 = \{11, 11.2, 13\}$, consistent with the inverse spectral-gap scaling predicted by Proposition 1. This magnitude analysis does not directly measure gradient direction; the theoretical decomposition of Section 4.2 establishes that the target-driven component T_{tgt} is directionally informative only when $\delta(\alpha, \Omega_h)$ is small. A prefactor initialized far from the target spectrum therefore receives a weaker and less directionally informative gradient signal, and training stalls before meaningful movement occurs.



(a) Prefactor displacement after 5,000 steps.

(b) Gradient magnitude vs. initial prefactor value.

Figure 2: Prefactor displacement and gradient analysis across 100 runs with $\alpha^{(0)} = \{1.01, 1.02, 1.03\}$. Displacement remains far below what is required to reach Ω_2 under all learning rates tested (a). Gradient magnitude is elevated near the target frequency range, consistent with the inverse spectral-gap scaling predicted by Proposition 1 (b).

4.2 Formal Characterization

We now prove that the gradient locality observed experimentally is a structural consequence of the circuit's Fourier architecture, holding for all ansatz parameter values θ and all learning rates.

Circuit partition. For each encoding gate j, partition the circuit as

$$U(x, \theta, \alpha) = U_{\text{after},j}(\theta_{\text{after}}, \alpha_{\setminus j}) \cdot S_j(x, \alpha_j) \cdot U_{\text{before},j}(\theta_{\text{before}}, \alpha_{\setminus j}), \tag{4}$$

where $U_{\text{before},j}$ and $U_{\text{after},j}$ collect all gates before and after gate j respectively, and $\alpha_{\setminus j}$ denotes all prefactors except α_j. Crucially, neither subcircuit depends on α_j itself. Define the pre-encoding state $|\psi_j(x, \theta)\rangle := U_{\text{before},j}|0\rangle$, the pulled-back observable $\widetilde{M}_j := U^\dagger_{\text{after},j} M U_{\text{after},j}$, and the effective observable $M_j^{\text{eff}} := S_j^\dagger \widetilde{M}_j S_j$.

The two lemmas below characterize the structure of the prefactor gradient; full proofs are given in Appendix C.

Assumption 1 (Symmetric input domain)**.** *Throughout this section, input data are assumed to lie in $x \in [-\pi, \pi]$, or equivalently to have been normalized to a symmetric interval around zero prior to encoding. This ensures $I(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} x\,\mathrm{d}x = 0$, which is used in the proof of Proposition 1. This assumption is without loss of generality: any bounded input domain $[a, b]$ can be mapped to $[-\pi, \pi]$ by the affine transformation $x \mapsto \frac{2\pi}{b-a}\left(x - \frac{a+b}{2}\right)$, which is absorbed into the encoding prefactors α_j.*

Lemma 1 (Commutator Identity)**.** *The gradient of the circuit output with respect to α_j satisfies*

$$\frac{\partial f}{\partial \alpha_j}(x) = ix \left\langle \psi_j \right| \left[M_j^{\text{eff}},\, H_j \right] \left| \psi_j \right\rangle \tag{5}$$

for all x, θ, and α.

The explicit factor of x in (5) distinguishes encoding parameter gradients from ansatz parameter gradients, where differentiation produces no such factor. As a non-periodic function on $[-\pi, \pi]$, x couples any Fourier mode $e^{i\nu x}$ to all other frequencies with amplitude $\sim 1/|\omega - \nu|$, so when the accessible spectrum $\Omega(\alpha)$ is far from the target support Ω_h, the coupling is weak.

Lemma 2 (Spectral Structure)**.** *Define $G_j(x) := \langle \psi_j | [M_j^{\text{eff}}, H_j] | \psi_j \rangle$, so that $\partial_{\alpha_j} f = ix\, G_j$ by Lemma 1. Then for all x, θ, and α:* (i) *G_j is a trigonometric polynomial with Fourier support $\Omega_j \subseteq \Omega(\alpha)$;* (ii) *$|G_j(x)| \leq 2\|M\|_{\text{op}} \|H_j\|_{\text{op}}$ uniformly in x and θ;* (iii) *each Fourier coefficient $d_\nu(\theta)$ of G_j satisfies $|d_\nu(\theta)| \leq 2\|M\|_{\text{op}} \|H_j\|_{\text{op}}$ uniformly in θ.*

Part (i) establishes that the probe function $\partial_{\alpha_j} f = ix\, G_j$ has spectral content confined to $\Omega(\alpha)$, regardless of θ. Parts (ii) and (iii) provide θ-free bounds on the magnitude of each contribution, which feed directly into the gradient suppression bound below.

Definition 4 (Spectral Interaction Gap)**.** *The* spectral interaction gap *between accessible spectrum $\Omega(\alpha)$ and target spectrum Ω_h is*

$$\delta(\alpha, \Omega_h) := \min_{\substack{\omega \in \Omega_h \\ \nu \in \Omega(\alpha)}} |\omega - \nu|. \tag{6}$$

Since $\Omega(\alpha)$ is symmetric (if $\nu \in \Omega(\alpha)$ then $-\nu \in \Omega(\alpha)$), we have $\min_{\nu \in \Omega(\alpha)} |\omega + \nu| = \min_{\nu \in \Omega(\alpha)} |\omega - \nu| = \delta(\alpha, \Omega_h)$ for any $\omega \in \Omega_h$, which is the form used in the proof of Proposition 1. For symmetric $\Omega(\alpha)$ and target Ω_h lying above the accessible range, $\delta(\alpha, \Omega_h) = \min_{\omega \in \Omega_h} \omega - \max \Omega(\alpha)$.

Proposition 1 (Spectral Gap Gradient Suppression)**.** *Decompose the loss gradient as $\partial_{\alpha_j} \mathcal{L} = T_{\text{tgt}} + T_{\text{self}}$, where $T_{\text{tgt}} = -\frac{1}{2\pi} \int h\, \partial_{\alpha_j} f \, \mathrm{d}x$ is the target-driven component and T_{self} the self-interaction. Provided $\delta(\alpha, \Omega_h) > 0$:*

$$|T_{\text{tgt}}| \leq \frac{4\, K(\alpha)\, \|M\|_{\text{op}}\, \|H_j\|_{\text{op}}}{\delta(\alpha, \Omega_h)} \sum_{\omega \in \Omega_h} |h_\omega|, \tag{7}$$

for all θ, α, and j.

Proof sketch. Substituting Lemma 1 and expanding in Fourier modes, T_{tgt} reduces to a sum of integrals $I(\omega + \nu) = \frac{1}{2\pi} \int_{-\pi}^{\pi} x\, e^{i(\omega+\nu)x}\, \mathrm{d}x$ over $(\omega, \nu) \in \Omega_h \times \Omega_j$. Integration by parts gives $|I(n)| \leq 2/|n|$ for all $n \neq 0$, and $I(0) = 0$ by oddness of x. Since $|\omega + \nu| \geq \delta(\alpha, \Omega_h)$ for symmetric $\Omega(\alpha)$, bounding each coefficient by Lemma 2(iii) and summing over $|\Omega_j| \leq K(\alpha)$ modes yields (7). The bound is uniform in θ because the coefficient bound in Lemma 2(iii) is. Full proof in Appendix C. □

Remark 1 (Self-interaction carries no target information)**.** *The self-interaction term T_{self} involves only cross-products of modes within $\Omega(\alpha)$ and contains no information about the location of Ω_h. Increasing $\|\theta\|$ enlarges $|T_{\text{self}}|$ but does not direct it toward the target spectrum. The suppression in Proposition 1 therefore cannot be overcome by modifying the ansatz initialization.*

Corollary 1 (Prefactor Displacement Bound)**.** *Under gradient descent with learning rate η for T steps,*

$$|\alpha_j^{(T)} - \alpha_j^{(0)}| \leq \eta T \cdot \frac{4\, K(\alpha^{(0)})\, \|M\|_{\text{op}}\, \|H_j\|_{\text{op}}}{\delta(\alpha^{(0)}, \Omega_h)} \sum_{\omega \in \Omega_h} |h_\omega| \;+\; \eta T \cdot 2\, K(\alpha^{(0)})^2\, \|M\|_{\text{op}}^2\, \|H_j\|_{\text{op}}. \tag{8}$$

For the experimental setup ($K = 7$ for a 3-qubit parallel circuit with prefactors near $\{1, 1, 1\}$, whose accessible spectrum $\{0, \pm 1, \pm 2, \pm 3\}$ has 7 elements; $\|M\|_{\text{op}} = 1$, $\|H_j\|_{\text{op}} = \frac{1}{2}$, $\sum |h_\omega| \approx 3$, $\delta = 8$, $\eta = 0.001$, $T = 5{,}000$), the target-driven term in (8) evaluates to ≈ 2.6, consistent with the empirically observed maximum displacement of $\approx$ 1–7 depending on learning rate (Figure 2(a)). The discrepancy reflects the conservatism of the supremum step; see Remark 2 below.

Remark 2 (Gradient descent vs. adaptive optimizers)**.** *The displacement bound* (8) *is stated for vanilla gradient descent and does not directly transfer to adaptive optimizers such as Adam (Kingma & Ba, 2017). A formal analog for Adam would require separate treatment accounting for the second-moment normalization schedule. The self-interaction term in* (8) *is also highly conservative in practice, as it bounds gradient magnitude uniformly over all training steps regardless of convergence behavior. Empirically, observed displacements are consistent with the target-driven term alone (Figure 2(a)), suggesting that the self-interaction contribution is substantially smaller than its worst-case bound in the regimes tested.*

5 Ternary Grid Initialization

5.1 Definition and Coverage

Definition 5 (Ternary Trainable-Frequency VQC)**.** *A ternary TF VQC initializes prefactors with exponentially-spaced values* $\alpha_j^{(0)} = 3^{j-1}$ *for* $j \in \{1, \ldots, k\}$*, which then become trainable alongside the ansatz parameters* θ*. For* k *encoding gates, the initial accessible spectrum is*

$$\Omega_{\text{grid}}^{(0)} = \left\{ n \in \mathbb{Z} : |n| \leq \tfrac{3^k-1}{2} \right\}, \tag{9}$$

covering all integers in $[-(3^k-1)/2, (3^k-1)/2]$ *(Shin et al., 2023; Peters & Schuld, 2023).*

Within the truncated Fourier series model of Section 3, the covering property of integer sets immediately implies that every target frequency $\omega^* \in [-\omega_{\max}, \omega_{\max}]$ lies within $\frac{1}{2}$ unit of some grid point, since the nearest integer to any real number is at most $\frac{1}{2}$ away. This holds independently of the random seed and any prior knowledge of which specific frequencies the target contains.

Proposition 2 (Ternary TF Parameter Requirements)**.** *A ternary TF circuit initialized to cover frequencies up to* $\omega_{\max}$ *requires* $k = \lceil \log_3(2\omega_{\max} + 1) \rceil$ *encoding gates and at least* $2\omega_{\max} + 1$ *ansatz parameters to independently control all Fourier coefficients accessible at initialization. As prefactors evolve during training, the accessible spectrum shifts accordingly; the parameter counts here refer to the initial configuration.*

Proof. The requirement $\omega_{\max} \leq (3^k - 1)/2$ gives $k \geq \log_3(2\omega_{\max} + 1)$, so $k = \lceil \log_3(2\omega_{\max} + 1) \rceil$ suffices. The ansatz parameter count follows from the fact that $|\Omega_{\text{grid}}^{(0)}| = 2\lfloor (3^k - 1)/2 \rfloor + 1 \geq 2\omega_{\max} + 1$ independent Fourier coefficients must be controllable, requiring at least that many free ansatz parameters (Schuld et al., 2021). □

Practical frequency specification. The approach requires specifying $\omega_{\max}$ before training. In practice this can be determined by progressively increasing k until validation performance saturates; the logarithmic scaling means this search requires only $\mathcal{O}(\log_3 \omega_{\max})$ trials. Alternatively, a spectral pre-analysis of the target values — for instance via a fast Fourier transform along one-dimensional slices of the input space — gives a practical estimate of the required frequency range per feature, though such marginal analyses may underestimate the full multivariate spectral content.

5.2 Formal Resolution of the Reachability Limitation

Ternary grid initialization simultaneously addresses the two aspects of the reachability barrier identified in Section 4: the large required prefactor displacement and the suppressed target-driven gradient signal.

Corollary 2 (Ternary Initialization Resolves Reachability)**.** *For ternary grid initialization* $\alpha_j^{(0)} = 3^{j-1}$ *with* $k = \lceil \log_3(2\omega_{\max} + 1) \rceil$ *encoding gates, every target frequency* $\omega^* \in [-\omega_{\max}, \omega_{\max}]$ *satisfies* $\delta(\alpha^{(0)}, \{\omega^*\}) \leq \frac{1}{2}$*. Substituting into* (7)*, the spectral gap suppression is removed:*

$$|T_{\text{tgt}}| \leq 8\, K(\alpha^{(0)})\, \|M\|_{\text{op}}\, \|H_j\|_{\text{op}} \sum_{\omega \in \Omega_h} |h_\omega|, \tag{10}$$

and the required prefactor displacement to reach any target within range is at most $\frac{1}{2}$ *unit. Together these ensure that the gradient signal is unsuppressed and the required movement is small at initialization, removing*

the structural obstruction to convergence. The convergence rate depends on circuit parameters and the learning rate schedule.

Proof. Since every real number is within $\frac{1}{2}$ of the nearest integer, and $\Omega^{(0)}_{\text{grid}}$ contains all integers up to $\omega_{\max}$ by construction with $k = \lceil \log_3(2\omega_{\max} + 1) \rceil$ encoding gates, every target frequency $\omega^* \in [-\omega_{\max}, \omega_{\max}]$ satisfies $\delta(\alpha^{(0)}, \{\omega^*\}) \leq \frac{1}{2}$. Substituting $\delta(\alpha^{(0)}, \Omega_h) \leq \frac{1}{2}$ into Proposition 1 immediately gives the stated gradient bound with factor $4/\delta \leq 8$, removing the $1/\delta$ penalty. The required prefactor displacement to reach any target within range is at most $\frac{1}{2}$ unit by the coverage property. Together, the unsuppressed gradient signal and the small required displacement remove the structural obstruction to convergence identified in Proposition 1; the convergence rate depends on circuit parameters and the learning rate schedule. □

Beyond resolving the spectral gap suppression, ternary initialization also inherits the near-uniform redundancy property established by Holzer & Turkalj (2024): at initialization, $R(\omega) \in \{1, 2\}$ across the accessible spectrum, avoiding the redundancy decay that further suppresses high-frequency gradient signals under unary encoding (Section 3.3).

The contrast with unary initialization is quantitative. For the experimental setting of Section 4 ($\Omega_2 = \{11, 11.2, 13\}$, $\alpha^{(0)} = \{1.01, 1.02, 1.03\}$), unary initialization faces $\delta = 8$, suppressing the target-driven gradient by a factor of $1/8$ while requiring ≈ 10 units of prefactor movement. Ternary initialization reduces the worst-case gap to $\delta \leq 0.2$ (the fractional distance from frequency 11.2 to its nearest grid point at 11), providing up to $40\times$ stronger gradient signal in this example while requiring at most 0.2 units of movement. In general, ternary initialization guarantees $\delta \leq \frac{1}{2}$, ensuring at least a $16\times$ improvement in gradient signal relative to the unary case here, jointly with a reduction in required displacement from ≈ 10 to $\leq \frac{1}{2}$ units.

Remark 3 (Two operational regimes)**.** *When the target has few frequencies relative to $\omega_{\max}$ (sparse regime: $k_{\text{opt}} \ll \log_3(2\omega_{\max} + 1)$), unary TF requires fewer encoding gates in principle. However, as Corollary 2 shows, this efficiency is unreachable in practice unless target frequencies happen to lie near the initialization. Ternary initialization trades a modest increase in encoding gate count for reliable convergence across the full target range. When target frequencies span the range or are unknown in advance (dense or agnostic regime), ternary initialization achieves both reliable convergence and an exponential reduction in encoding gate count relative to fixed unary approaches — the regime where its full advantage applies.*

5.3 Encoding Gate Complexity

Beyond resolving the reachability limitation, ternary initialization inherits the gate efficiency of fixed ternary encodings, which we now state formally.

Corollary 3 (Encoding Gate Complexity)**.** *A ternary TF circuit can implement any target single-qubit encoding $e^{i\omega_{\max} x H}$ to Frobenius norm precision ε — and thereby approximate target functions with frequency content up to $\omega_{\max}$ — using*

$$k \in \mathcal{O}\big(\log_3 \omega_{\max} + \log_3 \tfrac{1}{\varepsilon}\big) \tag{11}$$

encoding gates, compared to $\mathcal{O}(\omega_{\max} + \varepsilon^{-2})$ for unary encoding.

The proof proceeds via Fejér's theorem and Cesàro mean approximation and is given in Appendix B. The key steps are: ternary coverage of the integer component requires $k = \mathcal{O}(\log_3 \omega_{\max})$; fine precision is achieved by the Cesàro mean approximation error $\mathcal{O}(\sqrt{k} \cdot 3^{-k/2})$, giving $k = \mathcal{O}(\log_3(1/\varepsilon))$ for the precision requirement.

We note that while encoding gate count is reduced exponentially, total circuit complexity (encoding and ansatz gates combined) remains $\mathcal{O}(\omega_{\max})$ in both cases, since independently controlling all $2\omega_{\max} + 1$ accessible Fourier coefficients requires $\mathcal{O}(\omega_{\max})$ ansatz parameters regardless of the encoding strategy. The advantage of ternary encoding is specifically in the encoding gates, which are particularly resource-constrained on near-term quantum hardware.

6 Experiments

Having characterized the reachability limitation empirically and formally in Section 4, we now validate that ternary grid initialization resolves it across synthetic and real-world settings. All experiments use PennyLane (Bergholm et al., 2022) with JAX/JIT compilation (Frostig et al., 2018) and the Optax Adam optimizer (DeepMind, 2020) at learning rate $\eta = 0.001$ for 5,000 steps. Inputs are normalized to $[-\pi, \pi]$; outputs are scaled to $[-1, 1]$ via MinMaxScaler. We use an 80/20 train/test split and report R^2 on the held-out set, summarized as median and interquartile range (IQR) across 10 random weight initializations. Ansatz blocks use SpecialUnitary gates: full SU(2^n) blocks for the 3-qubit (63 parameters) and 4-qubit (255 parameters) circuits, and a hardware-efficient ansatz of 6 layers of 7 nearest-neighbour SU(4) blocks on 8 qubits with measurement on the central qubit (630 parameters) for the California Housing experiment. Four approaches are compared throughout: Fixed Unary, Trainable Unary ($\alpha^{(0)} = \{1.01, 1.02, 1.03\}$), Fixed Ternary, and Trainable Ternary ($\alpha^{(0)} = \{1, 3, 9\}$, ours).

6.1 Synthetic Benchmark

We construct target functions as truncated Fourier series with randomly sampled coefficients ($a_i, b_i, c_0 \sim$ Uniform($[0, 1)$)) at two frequency configurations. The first, $\Omega_1 = \{1, 1.2, 3\}$, replicates the original Jaderberg et al. (2024) setting. The second, $\Omega_2 = \{11, 11.2, 13\}$, shifts the entire spectrum by 10 units to isolate whether optimization can drive prefactors across a genuine spectral gap, using 100 equally spaced points over $[-\pi, \pi]$.

Results. Figure 1 showed all four approaches qualitatively; we now report the quantitative outcomes. On Ω_1 all approaches converge, confirming the architecture is expressive enough when target frequencies lie within the reachable range. On Ω_2, unary approaches fail as characterized in Section 4. Fixed Ternary achieves moderate performance, confirming that frequency coverage is important but not sufficient on its own — trainable prefactors provide additional fine-tuning that closes the gap between integer grid points and the non-integer target frequency 11.2. Trainable Ternary achieves median $R^2 = 0.9969$ with a minimum of 0.983; every one of the 100 runs reaches $R^2 > 0.95$.

Robustness across frequency offsets. Figure 3 shows performance as the frequency offset varies continuously from 0 to 9 units (spectra of the form $\Omega_1 + \mathcal{N}(\mu, 1)$ for $\mu \in [0, 9]$). Trainable Ternary maintains $R^2 > 0.95$ across all offsets. Unary approaches degrade monotonically with increasing shift. Fixed Ternary tracks closely with Trainable Ternary across most offsets, with Trainable Ternary providing consistent additional gain from prefactor fine-tuning.

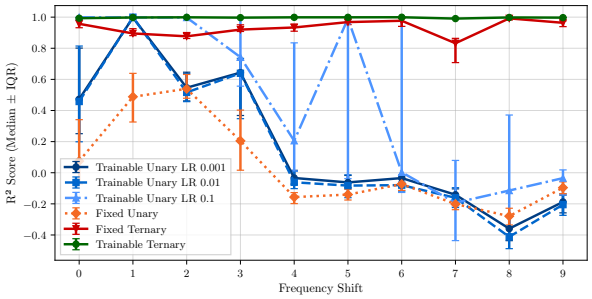


Figure 3: Median R^2 and IQR across 100 runs as the target frequency set shifts from Ω_1 to $\Omega_1 + 9$ (target coefficients $a_i, b_i, c_0 \sim$ Uniform($[0, 1)$)). Trainable Ternary maintains high performance across all offsets; unary approaches degrade monotonically with increasing shift.

Table 1: Comparison of optimization and initialization strategies on target $\Omega_2 = \{11, 11.2, 13\}$. Success is defined as $R^2 \geq 0.95$.

Method	Success	Median R^2
Unary init (Adam)	0%	−0.49
Random init (Adam)	39%	−0.002
CMA-ES (global)	25%	0.617
Ternary (ours)	**100%**	**0.988**

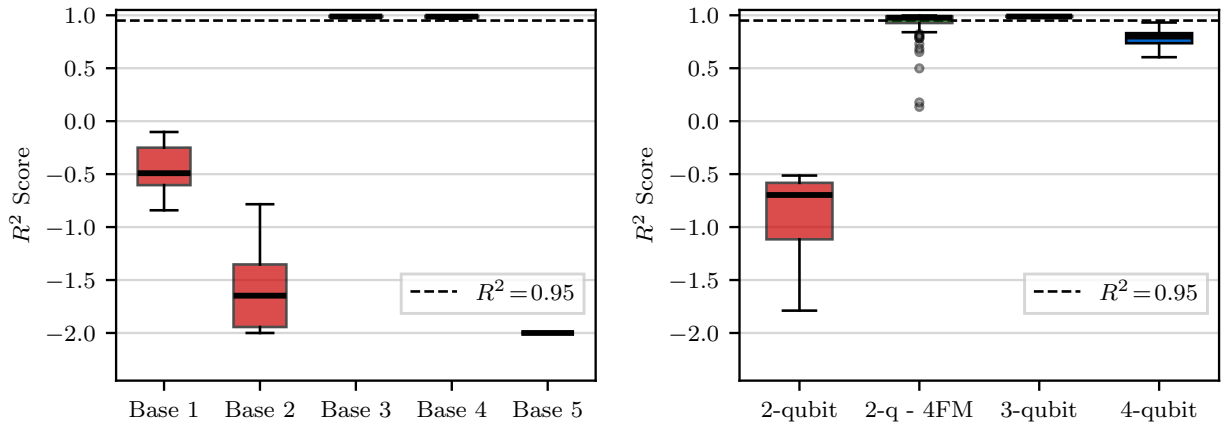


Figure 4: Ablation studies on $\Omega_2 = \{11, 11.2, 13\}$, 10 randomly chosen target coefficients with 10 different weight initialization seeds each. **(a)** R^2 distributions by encoding base, 3-qubit circuit with 1 feature map per qubit. Base-3 (ternary) and base-4 achieve 100% success; bases 1, 2, and 5 fail completely. **(b)** R^2 distributions by circuit architecture, ternary initialisation throughout. The 3-qubit (1 FM) and 2-qubit (2 FM) circuits achieve 100% success; the 4-qubit circuit is partially successful; the 2-qubit (1 FM) circuit fails. Dashed line marks $R^2 = 0.95$.

6.2 Ablation Studies

Encoding base and circuit architecture. Figure 4 examines two structural dimensions of ternary initialisation. Panel (a) varies the encoding base across $\{1, 2, 3, 4, 5\}$ on a 3-qubit circuit with target Ω_2. Only base-3 achieves reliable convergence across all 100 runs: base-2 fails because its accessible range $[-7, 7]$ does not reach the target; base-5 fails because, despite a wider range $[-31, 31]$, the target frequency 11 falls in one of its 36 integer gaps; base-4 succeeds on this particular target by coincidence — its spectrum happens to contain $\{11, 13\}$ — but has 16 gaps that would cause failure on any target falling within them.

Panel (b) fixes the ternary base and varies the circuit architecture. A 2-qubit circuit with 1 FM per qubit fails because its accessible range $[-4, 4]$ does not reach the target. A 3-qubit circuit with 1 FM per qubit succeeds as established above. A 4-qubit circuit with 1 FM per qubit covers the full target range $[-40, 40]$ but achieves only median $R^2 = 0.799$, confirming that frequency coverage is necessary but not sufficient on its own. Replacing the 4-qubit circuit with a 2-qubit circuit using 2 FMs per qubit — achieving the same accessible range $[-40, 40]$ with a different entanglement structure and 45 ansatz parameters — recovers 100% success, demonstrating that the underperformance of the 4-qubit circuit is attributable to circuit architecture rather than frequency coverage.

Taken together, both panels confirm the central claim: successful convergence requires that the initial prefactors place accessible frequencies close to the target spectrum. Panel (b) additionally shows that once coverage is ensured, circuit architecture has an independent effect on performance that warrants further investigation.

6.3 Alternative Optimization Strategies

The spectral gap suppression proved in Proposition 1 is a property of the gradient signal, leaving open whether a gradient-free optimizer or a different initialization strategy could navigate to the target frequencies from the standard unary starting point. We test both possibilities.

Derivative-free optimization: CMA-ES. We evaluate CMA-ES (Hansen, 2016) on the same 3-qubit architecture with target Ω_2, using a budget of 100,000 function evaluations (20× Adam's 5,000 steps), step size $\sigma_0 = 0.5$, and the same 100 runs. CMA-ES achieves $R^2 \geq 0.95$ in 25/100 runs (mean $R^2 = 0.983$ among

successes), compared to 0/100 for Adam with unary initialization. In the successful runs, prefactor values reach magnitudes up to ≈ 13, consistent with the target frequency, confirming that solutions exist and the loss surface is not flat near them. The remaining 75% of runs produce variable and often negative R^2 values, with high IQR. The low success rate despite $20\times$ the evaluation budget indicates that the reachability limitation is a property of the optimization landscape rather than of gradient-based optimization specifically: reaching the target from the standard initialization is difficult regardless of optimizer, and succeeds only when global search happens to land near the target spectrum.

Random prefactor initialization. A natural alternative to ternary initialization is to sample prefactors uniformly from a range spanning the target frequencies. We evaluate uniform initialization from Uniform($[1, 9]$), matching the range of the ternary grid $\{1, 3, 9\}$, on the same 100-run setup. The result is strongly bimodal: 39/100 runs achieve $R^2 \geq 0.95$ (mean 0.951 among successes), 50/100 produce negative R^2 (mean -0.87), and the remaining 11/100 fall in $[0, 0.95)$ (mean 0.20), reflecting runs where a randomly drawn prefactor happens to land near but not on the target spectrum. The median across all runs is $R^2 \approx -0.002$ with IQR spanning $[-1.05, +0.99]$. Inspection of successful runs reveals that they correspond to draws where the sampled prefactors happen to lie near the target frequencies; failures correspond to draws that do not. Random initialization therefore provides coverage in expectation but not by construction: whether any given run succeeds is entirely seed-dependent.

Table 1 places all approaches in context. Ternary grid initialization is the only strategy that achieves both reliable convergence and a principled frequency coverage property.

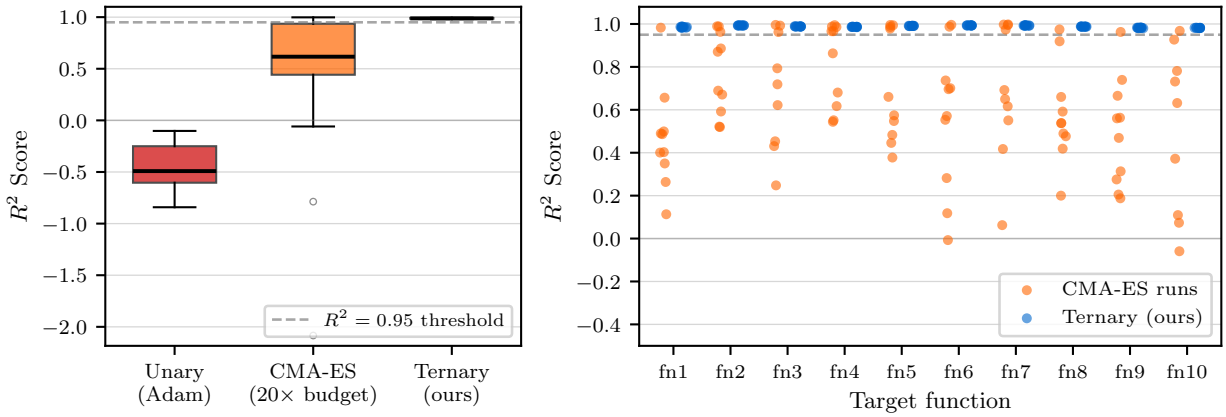


Figure 5: R^2 distributions across 100 runs for each optimization strategy on target Ω_2. **(a)** Overall distributions; percentages indicate success rates ($R^2 \geq 0.95$). **(b)** Per-function breakdown for CMA-ES and ternary across all 10 target functions: CMA-ES performance varies widely and unpredictably; ternary achieves tight, consistent $R^2 \approx 0.987$–0.993 for every function and seed.

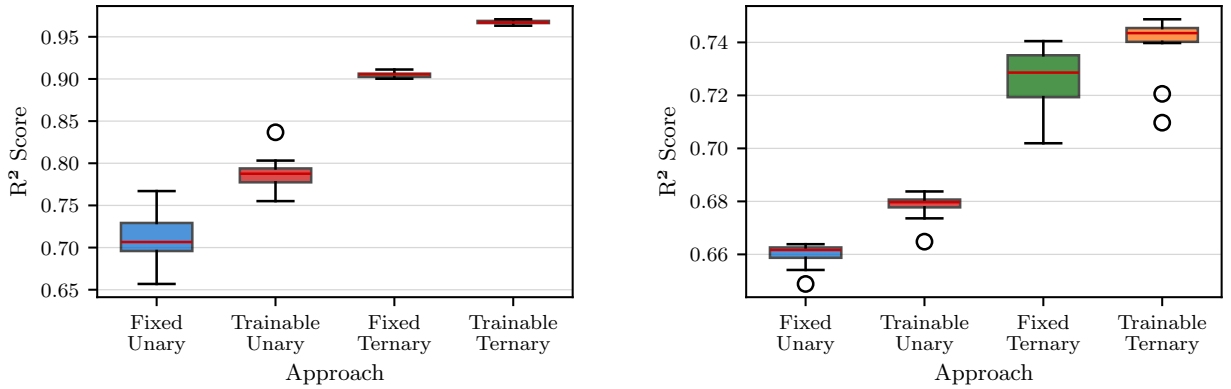
6.4 Real-World Validation: Flight Passengers

We validate on the Flight Passengers time series (Box & Jenkins, 1976) using a 4-qubit parallel architecture with one feature map per qubit (255 parameters).

Trainable Ternary achieves a 22.8% improvement over Trainable Unary. A Wilcoxon rank-sum test confirms all pairwise differences are statistically significant ($p = 0.000183$, $|r| = 1.0$ throughout), including Fixed vs. Trainable Ternary. The smaller advantage relative to the synthetic benchmark is expected: the synthetic experiment deliberately excluded frequencies below 10 to isolate long-range reachability, whereas real-world data contains spectral content across all frequencies including those already accessible under unary initialization.

Table 2: Median R^2 across real-world benchmarks. Flight Passengers uses a 4-qubit parallel circuit (255 parameters); California Housing uses an 8-qubit serial circuit with 6 layers of nearest-neighbour SU(4) gates and measurement on the central qubit (630 parameters).

Approach	Flight Passengers	California Housing
Fixed Unary	0.707	0.662
Trainable Unary	0.788	0.680
Fixed Ternary	0.906	0.730
Trainable Ternary (ours)	**0.967**	**0.744**



(a) Flight Passengers (4-qubit parallel, 1 feature map per qubit).

(b) California Housing (8-qubit serial, 5 feature maps per qubit).

Figure 6: R^2 distributions across 10 initializations on the two real-world benchmarks. The same performance staircase (Fixed Unary < Trainable Unary < Fixed Ternary < Trainable Ternary) appears on both datasets.

6.5 Multivariate Validation: California Housing

To confirm the approach extends to multivariate inputs, we evaluate on the California Housing dataset (Pace & Barry, 1997) with 8 input features, using an 8-qubit serial architecture with 5 encoding gates per qubit across 10 runs.

The four approaches produce the same performance staircase again (Figure 6b). Fixed Unary achieves median $R^2 = 0.662$ with a tight IQR $[0.659, 0.663]$: the deterministic integer-frequency spectrum gives every run near-identical coverage, so variance is minimal but the ceiling is low. Trainable Unary improves modestly to 0.680 through local prefactor adjustment within the reachable range. Fixed Ternary reaches 0.730 with substantially wider variance (IQR 0.700–0.740), reflecting sensitivity to non-integer target frequencies that fixed prefactors cannot adapt to. Trainable Ternary achieves both the highest median (0.744) and the tightest IQR among the ternary approaches (0.742–0.745). A Wilcoxon rank-sum test confirms that Trainable Ternary significantly outperforms Trainable Unary ($p = 0.000183$, $|r| = 1.0$). The Fixed vs. Trainable Ternary improvement is weaker here ($p = 0.014$, $|r| = 0.66$) than on Flight Passengers. California Housing is a high-dimensional regression task with complex spatial structure; the quantum model's frequency coverage, while improved by ternary initialization, remains limited relative to the full complexity of the target. The marginal benefit of prefactor fine-tuning is therefore smaller here than on Flight Passengers, where the dominant spectral structure is more directly addressable by the ternary grid.

7 Discussion and Limitations

A consistent staircase. Across all three datasets, the same performance ordering holds: Fixed Unary < Trainable Unary < Fixed Ternary < Trainable Ternary. This staircase is not coincidental: each step

corresponds to a distinct property identified by the theoretical analysis. The gap between Fixed and Trainable Unary reflects local prefactor fine-tuning within the narrow reachable range. The gap between Trainable Unary and Fixed Ternary reflects spectral gap suppression: even local fine-tuning cannot overcome a large $\delta(\alpha, \Omega_h)$, while dense integer coverage eliminates it at initialization. The gap between Fixed and Trainable Ternary reflects the adaptability advantage of learned prefactors over fixed ones in the presence of non-integer target frequencies.

The role of prefactor trainability. The comparison between Fixed and Trainable Ternary reveals a consistent pattern: trainability reduces variance and improves the median by using the integer grid points as initialization and fine-tuning within the locally reachable neighborhood to track actual target frequencies. In the limit of an infinitely dense fixed grid, the two approaches would converge in performance; in practice, with $k = \lceil \log_3(2\omega_{\max} + 1) \rceil$ encoding gates, the trainable variant achieves better adaptability at no additional circuit cost.

Encoding efficiency and prefactor coupling. One consequence of ternary encoding efficiency is that a small number of prefactors governs a large accessible spectrum: with $k = \lceil \log_3(2\omega_{\max} + 1) \rceil$ encoding gates, each prefactor simultaneously influences multiple frequencies. When target frequencies require movement in conflicting directions, a single prefactor cannot satisfy all constraints simultaneously — a coupling that does not arise in unary TF circuits to the same degree and which may limit fine-tuning on targets with complex spectral structure.

Gradient descent assumption. The theoretical results in Section 4.2 are stated for vanilla gradient descent. All experiments use Adam, whose second-moment normalization alters the effective step size in ways not captured by Corollary 1. Empirically, observed displacements are consistent with the target-driven component of the bound, suggesting the self-interaction term is substantially smaller than its worst-case value in the regimes tested. A formal extension of the displacement bound to adaptive optimizers would close this gap between theory and practice.

Continuous loss approximation. The theoretical results are stated for the continuous $L^2[-\pi, \pi]$ loss, whereas experiments minimize empirical mean squared error over a finite sample of 100 equally spaced points. For sufficiently dense, regularly spaced grids the two losses are close by standard quadrature arguments, but the precise gap between continuous and empirical guarantees is not characterized here and may matter for sparser or irregularly sampled data.

Tightening the displacement bound. Corollary 1 provides a rigorous upper bound on prefactor displacement; the target-driven component evaluates to ≈ 2.6 units at standard experimental parameters, consistent with the empirically observed limit, but the self-interaction term yields a highly conservative bound that does not reflect observed behavior. A tighter characterization that tracks the interaction between T_{self} and T_{tgt} over the course of training would sharpen the quantitative predictions of the theory, particularly for adaptive optimizers such as Adam.

Prefactor coupling and degrees of freedom. Ternary encoding achieves its gate efficiency precisely because each prefactor governs multiple frequencies simultaneously: k prefactors generate up to $(3^k - 1)/2$ accessible frequencies. For targets whose spectral structure aligns well with the ternary grid this constraint is benign; for targets requiring fine independent control of many frequencies simultaneously, the coupling between prefactors and accessible frequencies may limit the expressivity of the trained model relative to a unary TF circuit with more prefactor degrees of freedom. Characterizing which target structures are most affected, and whether hybrid encoding strategies can mitigate the coupling, is an open direction.

Encoding architecture and ansatz expressivity. The ablation in Section 6.2 shows that two circuits with identical accessible frequency ranges but different entanglement structures — a 4-qubit circuit with 1 FM per qubit and a 2-qubit circuit with 2 FMs per qubit — achieve substantially different performance. This demonstrates that frequency coverage alone does not determine convergence: circuit architecture has an

independent effect whose mechanism is not yet understood. Identifying which architectural properties govern performance beyond frequency coverage is left for future work.

Specification of $\omega_{\max}$. The ternary grid initialization requires specifying $\omega_{\max}$ in advance, which determines both the number of encoding gates and the initial prefactor values. In the absence of prior knowledge, a conservative over-estimate incurs only logarithmic overhead in encoding gates; a spectral pre-analysis of the target values via one-dimensional FFT slices per feature provides a more targeted estimate, though this may underestimate the full multivariate spectral content (Section 5). A fully automated method for inferring the required frequency range directly from training data would make the approach more broadly applicable.

Hardware-efficient ansatze. The 3-qubit and 4-qubit experiments use full SpecialUnitary blocks, which scale exponentially in parameter count and become impractical beyond small system sizes. The California Housing experiment already addresses this with a hardware-efficient nearest-neighbour arrangement of two-qubit SU(4) blocks (Wierichs et al., 2025). Whether the expressivity reduction that accompanies such hardware-efficient ansatze interacts with the frequency reachability properties established here — specifically, whether a restricted ansatz can independently control all Fourier coefficients accessible via ternary initialization — remains an open question for near-term experimental deployment.

Classical–quantum expressivity boundary. Trainable-frequency models incorporate learnable prefactors α_j that multiply the input data before quantum encoding, constituting a form of classical pre-processing. As Yu et al. (2022) acknowledge, this introduces ambiguity about whether expressivity improvements originate from the classical pre-processing or the quantum circuit itself. The ternary initialization strategy resolves a problem within the trainable-frequency framework as defined, but a precise characterization of which aspects of the modeling advantage require quantum resources and which are reproducible classically would substantially clarify the framework's scope.

Curse of dimensionality. For d-dimensional inputs, the number of accessible frequencies grows exponentially with d, and the parameter count required to control all Fourier coefficients scales accordingly (Pérez-Salinas et al., 2025). The California Housing experiment with 8 input features already requires a large ansatz, and the approach faces the same exponential scaling as classical Fourier methods in high dimensions. One direction toward mitigating this is selective frequency activation — learning which subset of accessible frequencies to use rather than controlling all of them — which reduces the effective dimensionality of the optimization problem and is explored in complementary work (Poppel et al., 2026).

8 Conclusion

Frequency reachability is a structural barrier in trainable-frequency quantum circuits: the prefactor gradient is suppressed by the spectral gap between the circuit's accessible frequencies and the target spectrum, independently of the ansatz, confining gradient-driven movement to a narrow neighborhood of initialization. Ternary grid initialization resolves this by construction, ensuring every target frequency within range lies within $\frac{1}{2}$ unit of a grid point and eliminating the gradient penalty simultaneously. The result is a principled initialization strategy that converts the theoretical encoding efficiency of trainable-frequency circuits into reliable empirical performance, confirmed by a consistent performance staircase across synthetic and real-world benchmarks.

The Fourier-analytic decomposition of the prefactor gradient into target-driven and self-interaction components provides a reusable analytical template for studying trainability in other parameterized encoding architectures. Open directions include extending the displacement bounds to adaptive optimizers, characterizing the prefactor coupling constraints that arise from ternary encoding efficiency, and addressing the exponential growth of the accessible frequency space in high-dimensional settings — the last of which selective frequency activation approaches may partially address.

References

Vasilis Belis, Joseph Bowles, Rishabh Gupta, Evan Peters, and Maria Schuld. Spectral methods: crucial for machine learning, natural for quantum computers?, 2026. URL https://arxiv.org/abs/2603.24654.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M. Sohaib Alam, Guillermo Alonso-Linaje, B. AkashNarayanan, Ali Asadi, Juan Miguel Arrazola, Utkarsh Azad, Sam Banning, Carsten Blank, Thomas R Bromley, Benjamin A. Cordier, Jack Ceroni, Alain Delgado, Olivia Di Matteo, Amintor Dusko, Tanya Garg, Diego Guala, Anthony Hayes, Ryan Hill, Aroosa Ijaz, Theodor Isacsson, David Ittah, Soran Jahangiri, Prateek Jain, Edward Jiang, Ankit Khandelwal, Korbinian Kottmann, Robert A. Lang, Christina Lee, Thomas Loke, Angus Lowe, Keri McKiernan, Johannes Jakob Meyer, J. A. Montañez-Barrera, Romain Moyard, Zeyue Niu, Lee James O'Riordan, Steven Oud, Ashish Panigrahi, Chae-Yeun Park, Daniel Polatajko, Nicolás Quesada, Chase Roberts, Nahum Sá, Isidor Schoch, Borun Shi, Shuli Shu, Sukin Sim, Arshpreet Singh, Ingrid Strandberg, Jay Soni, Antal Száva, Slimane Thabet, Rodrigo A. Vargas-Hernández, Trevor Vincent, Nicola Vitucci, Maurice Weber, David Wierichs, Roeland Wiersema, Moritz Willmann, Vincent Wong, Shaoming Zhang, and Nathan Killoran. Pennylane: Automatic differentiation of hybrid quantum-classical computations, 2022. URL https://arxiv.org/abs/1811.04968.

George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1976. Dataset: International airline passengers, monthly totals (1949-1960).

DeepMind. Optax: Gradient processing and optimization library in jax. https://optax.readthedocs.io/en/latest/api/optimizers.html, 2020. Accessed: 2025-07-29.

Callum Duffy and Marcin Jastrzebski. Spectral Bias in Variational Quantum Machine Learning. *arXiv preprint arXiv:2506.22555*, 6 2026.

Roy Frostig, Matthew Johnson, and Chris Leary. Compiling machine learning programs via high-level tracing. 2018. URL https://mlsys.org/Conferences/doc/2018/146.pdf.

Nikolaus Hansen. The CMA evolution strategy: A tutorial, 2016. URL https://arxiv.org/abs/1604.00772.

Patrick Holzer and Ivica Turkalj. Spectral invariance and maximality properties of the frequency spectrum of quantum neural networks, 2024. URL https://arxiv.org/abs/2402.14515.

Ben Jaderberg, Antonio A. Gentile, Youssef Achari Berrada, Elvira Shishenina, and Vincent E. Elfving. Let quantum neural networks choose their own frequencies. *Physical Review A*, 109(4), April 2024. ISSN 2469-9934. doi: 10.1103/physreva.109.042421. URL http://dx.doi.org/10.1103/PhysRevA.109.042421.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Martin Larocca, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J. Coles, Lukasz Cincio, Jarrod R. McClean, Zoë Holmes, and M. Cerezo. A review of barren plateaus in variational quantum computing, 2024. URL https://arxiv.org/abs/2405.00781.

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), November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07090-4. URL http://dx.doi.org/10.1038/s41467-018-07090-4.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3): 291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Adrián Pérez-Salinas, David López-Núñez, Artur García-Sáez, P. Forn-Díaz, and José I. Latorre. One qubit as a universal approximant. *Phys. Rev. A*, 104:012405, Jul 2021. doi: 10.1103/PhysRevA.104.012405. URL https://link.aps.org/doi/10.1103/PhysRevA.104.012405.

Evan Peters and Maria Schuld. Generalization despite overfitting in quantum machine learning models. *Quantum*, 7:1210, December 2023. ISSN 2521-327X. doi: 10.22331/q-2023-12-20-1210. URL http://dx.doi.org/10.22331/q-2023-12-20-1210.

Michael Poppel, David Bucher, Maximilian Zorn, Nico Kraus, Claudia Linnhoff-Popien, Philipp Altmann, and Jonas Stein. Mitigating exponential mixed frequency growth through frequency selection, 2026. URL https://arxiv.org/abs/2508.10533.

Adrián Pérez-Salinas, Mahtab Yaghubi Rad, Alice Barthe, and Vedran Dunjko. Universal approximation of continuous functions with minimal quantum circuits. *Physical Review Research*, 7(4), December 2025. ISSN 2643-1564. doi: 10.1103/t49h-jmty. URL http://dx.doi.org/10.1103/t49h-jmty.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3), March 2021. ISSN 2469-9934. doi: 10.1103/physreva.103.032430. URL http://dx.doi.org/10.1103/PhysRevA.103.032430.

S. Shin, Y. S. Teo, and H. Jeong. Exponential data encoding for quantum supervised learning. *Physical Review A*, 107(1), January 2023. ISSN 2469-9934. doi: 10.1103/physreva.107.012422. URL http://dx.doi.org/10.1103/PhysRevA.107.012422.

P. Turán. Sur les fonctions bornées et intégrables. In P. Turán (ed.), *Leopold Fejér Gesammelte Arbeiten I*, pp. 37–43. Birkhäuser, Basel, 1970.

Marco Wiedmann, Maniraman Periyasamy, and Daniel D. Scherer. Fourier analysis of variational quantum circuits for supervised learning, 2024. URL https://arxiv.org/abs/2411.03450.

David Wierichs, Korbinian Kottmann, and Nathan Killoran. Unitary synthesis with optimal brick wall circuits, 2025. URL https://arxiv.org/abs/2511.16736.

Zichao Yang, Alexander J Smola, Le Song, and Andrew Gordon Wilson. A la carte – learning fast kernels. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 1098–1106, 2015.

Zhan Yu, Hongshun Yao, Mujin Li, and Xin Wang. Power and limitations of single-qubit native quantum neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2751–2764, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b250de41980b58d34d6aadc3f4aedd4c-Paper-Conference.pdf.

Kaining Zhang, Liu Liu, Min-Hsiu Hsieh, and Dacheng Tao. Escaping from the barren plateau via Gaussian initializations in deep variational quantum circuits. *Physical Review Letters*, 134:150601, 2025. doi: 10.1103/PhysRevLett.134.150601.

Reproducibility Statement

Data and targets. Synthetic target functions are truncated Fourier series with randomly sampled coefficients $(a_i, b_i, c_0 \sim \text{Uniform}([0, 1)))$ at specified frequency spectra, evaluated at 100 equally spaced points over $[-\pi, \pi]$. Target outputs are scaled to $[-1, 1]$ via MinMaxScaler. The Flight Passengers dataset (Box & Jenkins, 1976) is publicly available (144 monthly observations); it is split sequentially 80/20 into train and test sets to respect temporal ordering, with values normalized to $[-1, 1]$ via MinMaxScaler. The California Housing dataset (Pace & Barry, 1997) is available via `scikit-learn` (Pedregosa et al., 2011), using an 80/20 random train/test split with the same output normalization.

Implementation. All experiments use PennyLane 0.42.0 with JAX/JIT compilation (JAX 0.5.0) and the Optax Adam optimizer at learning rate 0.001 for 5,000 steps. Mean squared error is the optimization objective; R^2 on an 80/20 train/test split is the evaluation metric. All training was performed on an Apple M2 MacBook Pro (16 GB RAM).

Ansatz initialization. All ansatz parameters are initialized by drawing uniformly from $[0, 2\pi]$ throughout, to isolate the effect of prefactor initialization from ansatz initialization effects.

Reporting. For synthetic benchmarks, results are reported as median and IQR across 10 random weight initializations per target function, with 10 target functions per configuration (100 runs total per approach). For real-world datasets, results are reported as median and IQR across 10 random weight initializations.

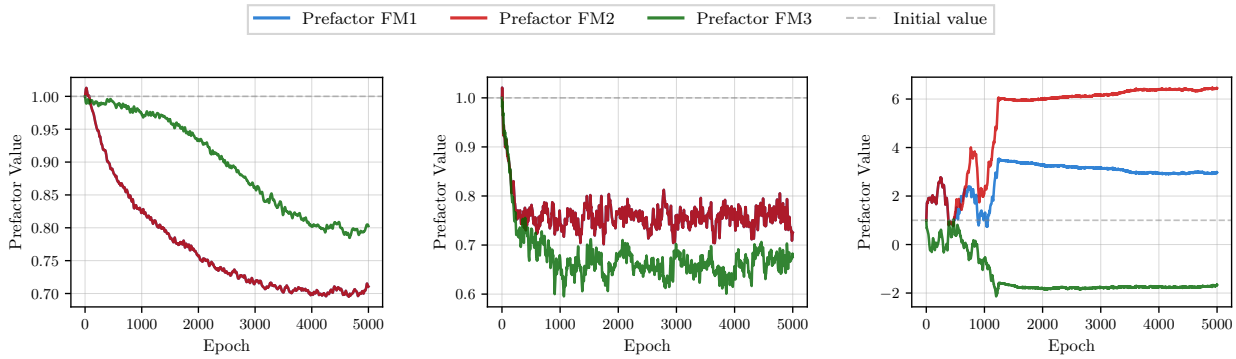
CMA-ES. CMA-ES experiments use the `cma` Python package with step size $\sigma_0 = 0.5$ and a budget of 100,000 function evaluations. Both ansatz weights and frequency prefactors are jointly optimized.

Circuit architectures. The 3-qubit synthetic circuit uses parallel $R_x(\alpha_i \cdot x)$ encoding on all qubits, interleaved with full 3-qubit SpecialUnitary ansatz blocks (63 parameters each); see Figure 10 in Appendix A. The 4-qubit Flight Passengers circuit uses the same parallel architecture with one feature map per qubit and full 4-qubit SpecialUnitary ansatz blocks (255 parameters each). The 8-qubit California Housing circuit uses one feature map per qubit with a hardware-efficient ansatz of 6 layers of 7 nearest-neighbour SU(4) blocks and measurement on qubit 3 (630 parameters).

Code. All experiment code and plotting notebooks are available at `https://github.com/mpoppel/frequency_reachability`.

A Supporting Experimental Figures

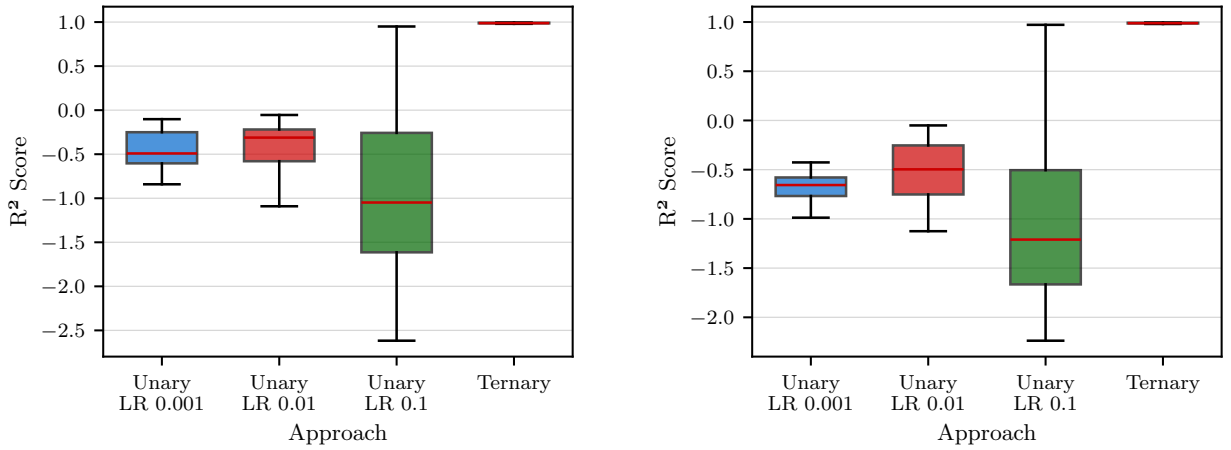
The following figures support the robustness claims in Section 4.1. Figure 7 shows prefactor evolution under identical initialization $\{1.0, 1.0, 1.0\}$, confirming that lock-step symmetry does not prevent failure. Figure 8 confirms that the failure mode is insensitive to whether prefactors are initialized identically or with a small perturbation.



(a) $\eta = 0.001$ (b) $\eta = 0.01$ (c) $\eta = 0.1$

Figure 7: Prefactor evolution under identical initialization $\{1.0, 1.0, 1.0\}$ for three learning rates. At $\eta \leq 0.01$ prefactors evolve lock-step; at $\eta = 0.1$ symmetry breaks but most prefactors still fail to reach Ω_2.

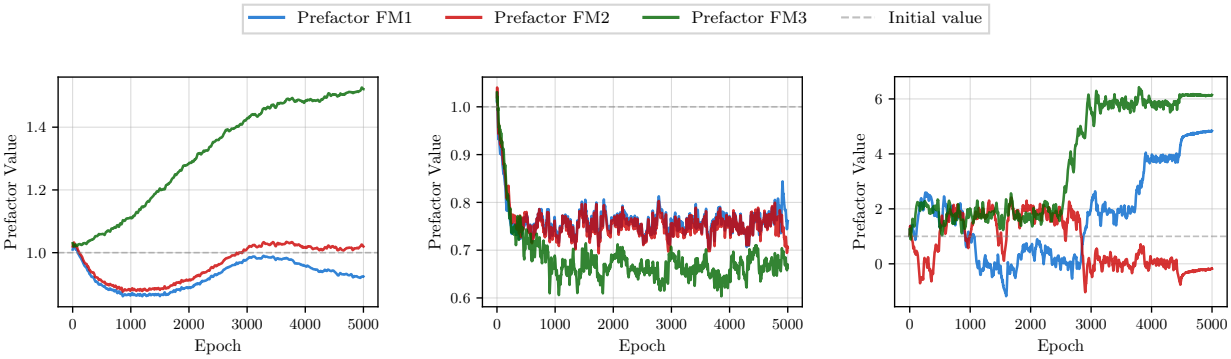
Figure 9 shows prefactor trajectories for the perturbed initialization $\{1.01, 1.02, 1.03\}$ across three learning rates, confirming that only $\eta = 0.1$ enables substantial movement but without reliable convergence.



(a) Identical init $\{1.0, 1.0, 1.0\}$.

(b) Perturbed init $\{1.01, 1.02, 1.03\}$.

Figure 8: R^2 distributions across 100 runs for target $\Omega_2 = \{11, 11.2, 13\}$, comparing identical and perturbed unary initializations against trainable ternary ($\eta = 0.001$). Both unary strategies produce similar failure distributions despite different prefactor evolution patterns.



(a) $\eta = 0.001$

(b) $\eta = 0.01$

(c) $\eta = 0.1$

Figure 9: Prefactor evolution for unary initialization on $\Omega_2 = \{11, 11.2, 13\}$, showing the best-performing run at each learning rate. Only $\eta = 0.1$ enables substantial movement, but convergence remains unreliable at that rate (Figure 8).

B Proof of Ternary Encoding Gate Complexity

We prove Corollary 3, extending Corollary 2 of Pérez-Salinas et al. (2025) to the ternary encoding setting. Throughout this appendix we use the normalized domain $x \in [0, 1]$ and Fourier bases $e^{in\pi x}$ for consistency with the cited proof; all results transfer to $x \in [-\pi, \pi]$ by rescaling.

B.1 Circuit Architecture

Definition 6 (Ternary Encoding Circuit)**.** *A ternary encoding circuit with k encoding gates follows the WSW (Weight-Signal-Weight) architecture:*

$$U_k^{\text{ternary}}(\theta, x) = W_k(\theta_k) \cdot S_{k-1}(x) \cdot W_{k-1}(\theta_{k-1}) \cdots S_1(x) \cdot W_1(\theta_1) \cdot S_0(x) \cdot W_0(\theta_0), \tag{12}$$

where $S_j(x) = e^{i \cdot 3^j \pi x \sigma_z}$ are fixed encoding gates with exponentially-spaced prefactors, $W_j(\theta_j)$ are parametrized ansatz blocks, and $x \in [0, 1]$.

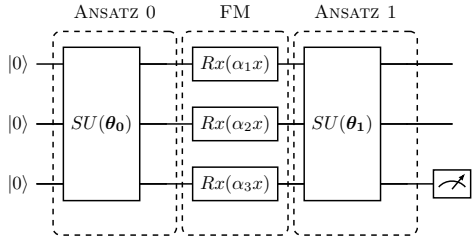


Figure 10: 3-qubit circuit with parallel $R_x(\alpha_i x)$ encoding and SpecialUnitary ansatz blocks (63 trainable parameters each).

By Shin et al. (2023); Peters & Schuld (2023), k ternary gates provide access to all integer frequencies in $\Omega_k = \{n \in \mathbb{Z} : |n| \leq M\}$ where $M = (3^k - 1)/2$.

B.2 Mathematical Background

Theorem 3 (Fejér's Theorem, Turán 1970)**.** *Let $f \in C([0, 2])$ be a continuous 2-periodic function with partial Fourier sums $S_N(f)(x)$. The Cesàro means $\sigma_N(f)(x) = \frac{1}{N+1} \sum_{k=0}^{N} S_k(f)(x)$ converge uniformly to f on $[0, 2]$.*

By the generalized quantum signal processing framework of Pérez-Salinas et al. (2025), any polynomial $P(e^{ix})$ of degree at most M can be implemented in the circuit matrix; in particular, $P_M(x)$ can be chosen as the Cesàro mean of e^{iwx}.

B.3 Proof of Corollary 3

Proof. **Step 1 (Coarse frequency coverage).** Any target frequency $w \in \mathbb{R}$ decomposes as $w = K\pi + w'$ with $|w'| \leq \pi/2$. Ternary coverage of the integer component requires $M \geq K$, giving $k \geq \log_3(2w/\pi + 1) = \mathcal{O}(\log_3 w)$.

Step 2 (Fine precision). Define the auxiliary function $a(x) = e^{iwx}$ for $x \in [0, 1]$ and $a(x) = e^{iw(2-x)}$ for $x \in (1, 2)$. This function is continuous and 2-periodic with Fourier coefficients $c_n = -\frac{i}{w} \frac{((-1)^n e^{iw} - 1)}{w^2 - (n\pi)^2}$. Constructing the Cesàro mean $P_M(x)$ over all accessible frequencies, the supremum approximation error satisfies

$$\sup_{x \in [0,1]} |P_M(x) - e^{iwx}| \in \mathcal{O}(M^{-1} \log M) = \mathcal{O}(k \cdot 3^{-k}). \tag{13}$$

The Frobenius norm error at the matrix level is therefore $\sup_x \|R_M - e^{iwx\sigma_z}\|_F \in \mathcal{O}(\sqrt{k} \cdot 3^{-k/2})$.

Step 3 (Precision requirement). Setting $C\sqrt{k} \cdot 3^{-k/2} \leq \varepsilon$ gives $k = \mathcal{O}(\log_3(1/\varepsilon))$, absorbing doubly-logarithmic factors.

Step 4 (Total). Combining Steps 1 and 3: $k \geq \max\{\mathcal{O}(\log_3 w), \mathcal{O}(\log_3(1/\varepsilon))\} = \mathcal{O}(\log_3 w + \log_3(1/\varepsilon))$. □

Remark 4 (Ansatz parameter requirements)**.** *While $k = \mathcal{O}(\log_3 \omega_{\max})$ encoding gates suffice, independently controlling all Fourier coefficients over the accessible range $[-M, M]$ requires $|\theta| = \mathcal{O}(M) = \mathcal{O}(\omega_{\max})$ ansatz parameters (Schuld et al., 2021). Total circuit complexity is therefore $\mathcal{O}(\omega_{\max})$ for both unary and ternary approaches; the advantage of ternary encoding lies specifically in the encoding gate count.*

C Proofs of Main Theoretical Results

This appendix provides full proofs of the results stated in Section 4.2. We restate each result before its proof for readability.

Circuit partition and notation. For each encoding gate j, partition the circuit as

$$U(x,\theta,\alpha) = U_{\text{after},j}(\theta_{\text{after}},\alpha_{\setminus j}) \cdot S_j(x,\alpha_j) \cdot U_{\text{before},j}(\theta_{\text{before}},\alpha_{\setminus j}), \tag{14}$$

where $\alpha_{\setminus j}$ denotes all prefactors except α_j, and neither $U_{\text{before},j}$ nor $U_{\text{after},j}$ depends on α_j. Recall the definitions of the pre-encoding state $|\psi_j(x,\theta)\rangle := U_{\text{before},j}|0\rangle$, the pulled-back observable $\widetilde{M}_j := U_{\text{after},j}^\dagger M U_{\text{after},j}$, and the effective observable $M_j^{\text{eff}} := S_j^\dagger \widetilde{M}_j S_j$ from Section 4.2.

C.1 Proof of Lemma 1 (Commutator Identity)

Proof. Using partition (14), write $f(x) = \langle\psi_j|\, S_j^\dagger\, \widetilde{M}_j\, S_j\, |\psi_j\rangle$. Since neither $|\psi_j\rangle$ nor $\widetilde{M}_j$ depends on α_j, the product rule gives

$$\frac{\partial f}{\partial \alpha_j} = \Big\langle \psi_j \;\Big|\; \frac{\partial S_j^\dagger}{\partial \alpha_j}\, \widetilde{M}_j\, S_j \;\Big|\; \psi_j \Big\rangle + \Big\langle \psi_j \;\Big|\; S_j^\dagger\, \widetilde{M}_j\, \frac{\partial S_j}{\partial \alpha_j} \;\Big|\; \psi_j \Big\rangle. \tag{15}$$

Direct differentiation of $S_j = e^{i\alpha_j x H_j}$ yields $\partial_{\alpha_j} S_j = ixH_jS_j$ and $\partial_{\alpha_j} S_j^\dagger = -ixS_j^\dagger H_j$. Substituting:

$$\frac{\partial f}{\partial \alpha_j} = ix\,\langle\psi_j|\, S_j^\dagger(-H_j\widetilde{M}_j + \widetilde{M}_jH_j)S_j\,|\psi_j\rangle = ix\,\langle\psi_j|\, S_j^\dagger[\widetilde{M}_j,\, H_j]S_j\,|\psi_j\rangle. \tag{16}$$

Since $[H_j, S_j] = 0$ (the gate commutes with its own generator), $S_j^\dagger H_j S_j = H_j$, so $S_j^\dagger[\widetilde{M}_j, H_j]S_j = [S_j^\dagger \widetilde{M}_j S_j,\, H_j] = [M_j^{\text{eff}},\, H_j]$, completing the proof. □

C.2 Proof of Lemma 2 (Spectral Structure)

Proof. Part (i). Both $|\psi_j(x,\theta)\rangle$ and $\widetilde{M}_j(x,\theta)$ depend on x through the encoding gates indexed by $\alpha_{\setminus j}$, contributing Fourier factors $e^{\pm i\alpha_\ell x H_\ell}$ for $\ell \neq j$. The effective observable $M_j^{\text{eff}} = S_j^\dagger \widetilde{M}_j S_j$ acquires an additional conjugation by $e^{\pm i\alpha_j x H_j}$. All matrix elements of M_j^{eff} are therefore finite linear combinations of terms $e^{i\omega x}$ with $\omega \in \Omega(\alpha)$, and this inclusion is inherited by $[M_j^{\text{eff}}, H_j]$ and by the scalar expectation value $G_j(x)$, giving $\Omega_j \subseteq \Omega(\alpha)$.

Part (ii). The commutator satisfies $\|[M_j^{\text{eff}}, H_j]\|_{\text{op}} \le 2\|M_j^{\text{eff}}\|_{\text{op}}\|H_j\|_{\text{op}}$. Since conjugation by a unitary preserves operator norms, $\|M_j^{\text{eff}}\|_{\text{op}} = \|M\|_{\text{op}}$. Applying to the unit-vector expectation value: $|G_j(x)| \le \|[M_j^{\text{eff}}, H_j]\|_{\text{op}} \le 2\|M\|_{\text{op}}\|H_j\|_{\text{op}}$, uniformly in x, θ, and α.

Part (iii). Each Fourier coefficient satisfies $|d_\nu(\theta)| = \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} G_j(x)e^{-i\nu x}\,\mathrm{d}x\right| \le \sup_x |G_j(x)| \le 2\|M\|_{\text{op}}\|H_j\|_{\text{op}}$, uniformly in θ. □

C.3 Auxiliary Bound

Lemma 3 (Operator Norm Bound)**.** *For any α, θ:*

$$|\partial_{\alpha_j}\mathcal{L}| \le 2\pi\|M\|_{\text{op}}\|H_j\|_{\text{op}}\|f-h\|_{L^2}. \tag{17}$$

Proof. Writing the gradient as an $L^2([-\pi,\pi])$ inner product,

$$\frac{\partial \mathcal{L}}{\partial \alpha_j} = \langle f-h,\, \partial_{\alpha_j} f\rangle_{L^2}. \tag{18}$$

Cauchy–Schwarz gives $|\partial_{\alpha_j}\mathcal{L}| \le \|f-h\|_{L^2}\|\partial_{\alpha_j} f\|_{L^2}$. By Lemma 1 and Lemma 2(ii), $|\partial_{\alpha_j} f(x)| = |x||G_j(x)| \le \pi \cdot 2\|M\|_{\text{op}}\|H_j\|_{\text{op}}$ for all $x \in [-\pi,\pi]$, so $\|\partial_{\alpha_j} f\|_{L^2} \le 2\pi\|M\|_{\text{op}}\|H_j\|_{\text{op}}$. Since $\mathcal{L} = \|f-h\|_{L^2}^2$, the result follows. □

Remark 5. *Lemma 3 provides a baseline bound that does not capture the spectral gap structure. Proposition 1 gives a sharper characterization by expanding the gradient in Fourier modes and applying the gap $\delta(\alpha,\Omega_h)$ explicitly.*

C.4 Proof of Proposition 1 (Spectral Gap Gradient Suppression)

Proof. Substituting Lemma 1 and expanding in Fourier modes $h(x) = \sum_{\omega\in\Omega_h} h_\omega e^{i\omega x}$ and $G_j(x) = \sum_{\nu\in\Omega_j} d_\nu e^{i\nu x}$:

$$T_{\text{tgt}} = -i\sum_{\substack{\omega\in\Omega_h\\ \nu\in\Omega_j}} h_\omega d_\nu \underbrace{\frac{1}{2\pi}\int_{-\pi}^{\pi} x\, e^{i(\omega+\nu)x}\,\mathrm{d}x}_{=:\, I(\omega+\nu)}. \tag{19}$$

Step 1: Evaluate $I(n)$. For $n \in \mathbb{R}\setminus\{0\}$, integration by parts gives

$$I(n) = \frac{\cos(n\pi)}{in} - \frac{\sin(n\pi)}{\pi n^2}, \qquad |I(n)| \le \frac{1}{|n|} + \frac{1}{\pi|n|} \le \frac{2}{|n|}. \tag{20}$$

For $n = 0$: $I(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} x\,\mathrm{d}x = 0$ by oddness of x on $[-\pi,\pi]$ (Assumption 1), so $n = 0$ terms contribute nothing to (19).

Step 2: Apply the spectral gap. For all pairs with $\omega + \nu \neq 0$, note that the accessible spectrum $\Omega(\alpha)$ is symmetric: if $\nu \in \Omega(\alpha)$ then $-\nu \in \Omega(\alpha)$. Therefore $|\omega + \nu| = |\omega - (-\nu)|$, and since $-\nu \in \Omega(\alpha)$, this equals the distance from $\omega \in \Omega_h$ to an element of $\Omega(\alpha)$, which is at least $\delta(\alpha, \Omega_h)$ by definition. Substituting (20) into (19):

$$|T_{\text{tgt}}| \le \frac{2}{\delta(\alpha,\Omega_h)} \sum_{\omega\in\Omega_h} |h_\omega| \cdot \sum_{\nu\in\Omega_j} |d_\nu(\theta)|. \tag{21}$$

Step 3: Bound $\sum_\nu |d_\nu|$. By Lemma 2(iii), $|d_\nu(\theta)| \le 2\|M\|_{\text{op}}\|H_j\|_{\text{op}}$ uniformly in θ. With $|\Omega_j| \le K(\alpha)$:

$$\sum_{\nu\in\Omega_j} |d_\nu(\theta)| \le K(\alpha)\cdot 2\|M\|_{\text{op}}\|H_j\|_{\text{op}}. \tag{22}$$

Substituting into (21) yields the bound. Uniformity in θ follows from Lemma 2(iii). □

Remark 6 (Self-interaction carries no target information)**.** *The self-interaction term $T_{\text{self}} = \frac{1}{2\pi}\int f\,\partial_{\alpha_j} f\,\mathrm{d}x$ involves only cross-products of modes within $\Omega(\alpha)$ and contains no information about the location of Ω_h. Applying* (20) *with the minimum nonzero intra-spectrum cross-sum $|\omega + \nu| \ge 1$ for integer-frequency circuits gives $|T_{\text{self}}| \le 2K(\alpha)^2\|M\|_{\text{op}}^2\|H_j\|_{\text{op}}$, a constant independent of $\delta(\alpha,\Omega_h)$. Modifying the ansatz initialization therefore cannot overcome the spectral gap suppression in Proposition 1.*

C.5 Proof of Corollary 1 (Prefactor Displacement Bound)

Proof. By the triangle inequality,

$$|\alpha_j^{(T)} - \alpha_j^{(0)}| \le \sum_{t=0}^{T-1} \eta \left|\frac{\partial\mathcal{L}}{\partial\alpha_j}\right|^{(t)} \le \eta T \cdot \sup_t \left|\frac{\partial\mathcal{L}}{\partial\alpha_j}\right|^{(t)}. \tag{23}$$

Bounding the gradient at each step by the decomposition $|\partial_{\alpha_j}\mathcal{L}| \le |T_{\text{tgt}}| + |T_{\text{self}}|$, applying Proposition 1 to $|T_{\text{tgt}}|$ and the self-interaction bound to $|T_{\text{self}}|$, and evaluating at $\alpha^{(0)}$ (self-consistent in the failure regime: since $\delta(\alpha^{(0)},\Omega_h)$ is large, the target-driven gradient is suppressed, prefactors move little, and δ remains close to its initial value throughout training) yields the stated bound. □