Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Training-Free Certified Bounds for Quantum Regression: A Scalable Framework

Demerson N. Gonçalves

demerson.goncalves@cefet-rj.br

Federal Center for Technological Education Celso Suckow da Fonseca

Tharso D. Fernandes

Universidade Federal do Espírito Santo

Pedro H. G. Lugao

Federal Center for Technological Education Celso Suckow da Fonseca

João T. Dias

Federal Center for Technological Education Celso Suckow da Fonseca

Research Article

Keywords: Quantum kernel regression, Pauli decomposition, Monte Carlo sampling, Feature map evaluation

Posted Date: March 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8660030/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Training-Free Certified Bounds for Quantum Regression: A Scalable Framework

Demerson N. Gonçalves[1*], Tharso D. Fernandes[2], Pedro H. G. Lugao[3], João T. Dias[4]

[1*]Dept. of Mathematics, Federal Center for Technological Education Celso Suckow da Fonseca (CEFET-RJ), Petrópolis, RJ, Brazil.
[2]Dept. of Mathematics, Federal University of Espírito Santo (UFES), Alegre, ES, Brazil.
[3]Dept. of Computer Engineering, CEFET-RJ, Petrópolis, RJ, Brazil.
[4]Dept. of Telecommunications Engineering, CEFET-RJ, Rio de Janeiro, RJ, Brazil.

*Corresponding author(s). E-mail(s): demerson.goncalves@cefet-rj.br;
Contributing authors: tharso.fernandes@ufes.br;
pedro.lugao@cefet-rj.br; joao.dias@cefet-rj.br;

Abstract

We present a training-free, certified error bound for quantum regression derived directly from Pauli expectation values. Generalizing the heuristic of minimum accuracy from classification to regression, we evaluate axis-aligned predictors within the Pauli feature space. We formally prove that the optimal axis-aligned predictor yields a rigorous upper bound on the minimum training Mean Squared Error (MSE) attainable by any linear or kernel-based regressor defined on the same quantum feature map. Since computing this bound exactly requires an intractable scan over the full Pauli basis, we introduce an adaptive Monte Carlo framework that estimates it from a tractable subset of axes and provides non-asymptotic statistical guarantees under a finite measurement budget. The resulting procedure enables rapid, training-free comparison of quantum feature maps and early diagnosis of feature-map capacity through the observed density of high-quality Pauli axes, supporting informed architecture selection prior to deploying higher-complexity models.

Keywords: Quantum kernel regression, Pauli decomposition, Monte Carlo sampling, Feature map evaluation.

1 Introduction

The capacity of machine learning (ML) to extract actionable patterns from high-dimensional data has established it as a standard paradigm for decision-making under uncertainty and automated discovery in complex systems [1, 2]. Quantum machine learning (QML) is a rapidly growing field that aims to transcend classical ML limitations by exploiting the unique structure and computational resources of quantum systems [3, 4]. One of the most promising QML paradigms relies on quantum kernel methods. Central to this framework is the *feature map*, a transformation that embeds classical input data into a high-dimensional Hilbert space to render complex, non-linear relationships amenable to linear analysis. In the quantum setting, this map encodes data into quantum states via parameterized quantum circuits, allowing classical algorithms to operate on the resulting quantum-induced feature space. This hybrid quantum–classical framework was established in the foundational works of Havlíček et al. [5], Schuld and Killoran [6], and Mitarai et al. [7]. Building on these seminal contributions, subsequent research has extensively investigated the hardware implementation and theoretical properties of these maps [8–11].

However, leveraging these quantum feature spaces in practice involves significant hurdles. In variational approaches, the optimization of parameterized circuits is frequently obstructed by the phenomenon of *barren plateaus*, where the gradients of the cost function vanish exponentially with the number of qubits, rendering deep architecture training effectively impossible [12–14]. To navigate this landscape without full training, geometric descriptors such as *expressibility* and *entangling capability* [15] have been proposed to characterize the ansatz structure. While these task-agnostic metrics provide valuable insights into the statistical distribution of states, they do not directly predict the generalization performance or the training error for a specific target function. Furthermore, although quantum kernel methods circumvent gradient-based optimization, they face severe scalability constraints in the Noisy Intermediate-Scale Quantum (NISQ) era [16], where limited coherence times and hardware noise make the estimation of full kernel matrices $N \times N$ prohibitively expensive. Validating a quantum model also requires benchmarking against robust classical baselines, such as standard Support Vector Regression (SVR) [17], which sets a high threshold for performance.

Consequently, assessing and comparing quantum feature maps by fully training models for every candidate architecture becomes computationally intractable. In kernel-based regression, this typically necessitates constructing the full kernel matrix and tuning regularization hyperparameters; in variational approaches, it requires iterative circuit evaluations for gradient updates. These costs scale poorly with dataset size and circuit depth. To address similar challenges in classification, Suzuki et al. [18] proposed the *minimum accuracy* heuristic. This method estimates the baseline performance obtainable by restricting measurements to axis-aligned Pauli observables, avoiding explicit optimization. However, the original formulation focuses exclusively on classification and relies on a full scan of the Pauli basis. This creates two significant gaps: first, there is no direct analogue for regression tasks, where minimizing continuous error, such as the mean squared error (MSE), is paramount; second, the computational cost of a full Pauli decomposition scales as 4^n for n qubits, rendering the exact metric intractable for larger systems.

In this work, we extend the analysis and practical utility of feature map diagnostics in three main directions, generalizing the axis-aligned concept to quantum regression. First, we formally define a regression setting in the Pauli feature space and introduce a training-free score based on single-coordinate least squares. Unlike geometric heuristics, this metric is directly grounded in the minimization of the MSE. Second, we provide a rigorous theoretical result (Theorem 1) proving that the MSE of the best axis-aligned predictor constitutes a certified *upper bound* on the minimum training MSE attainable by any linear or kernel-based regressor on the same feature map. This formally justifies the score as a "guaranteed limit": if the axis-aligned error is low, the full quantum model is mathematically guaranteed to perform at least as well. Third, we introduce a statistically-certified Monte Carlo (MC) framework to render this bound computationally viable. Instead of the intractable 4^n scan, we introduce an estimator, $\widehat{\mathrm{MSE}}_{\mathrm{axis}}$, computed from a random subset of axes. We derive non-asymptotic probability guarantees based on Hoeffding's concentration inequalities [19], relating the sample size to the reliability of the bound (Theorem 3).

From a practical standpoint, our results transform this theoretical bound into a scalable tool for pre-screening quantum feature maps. The proposed workflow allows one to avoid the exponential cost of characterizing the full feature space. By drawing a random sample of $t \ll 4^n$ Pauli axes and computing simple 1D regression scores, one obtains a certified upper bound on the model's potential error. This enables the rapid comparison of architectures, early diagnosis of expressivity issues, and informed resource allocation, effectively filtering out poor feature maps before deploying higher-complexity models. Consequently, the computational burden of feature map selection is drastically reduced from the polynomial complexity of full kernel training (typically $\mathcal{O}(N^3)$) to a scalable cost dominated by the number of sampled axes t, where t can be kept small while maintaining rigorous statistical confidence.

The remainder of this paper is organized as follows. Section 2 establishes the theoretical framework, formally defining the axis-aligned regression score and proving it constitutes a certified upper bound on the optimal training error. Section 3 introduces the Monte Carlo estimation strategy designed to overcome the exponential scaling of the feature space; this section also derives rigorous non-asymptotic statistical guarantees and details the adaptive algorithm for sample-size calibration. The experimental methodology, including dataset generation and feature map configurations, is detailed in Section 4. Section 5 presents the numerical benchmarks, validating the estimator's tightness and predictive power against fully trained quantum and classical regressors. Finally, Section 6 concludes the work and outlines directions for future research.

2 Pauli-axis Upper Bound for Quantum Regression

We consider supervised regression in the quantum kernel framework, where the goal is to predict a real-valued label $y \in \mathbb{R}$ from an input $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^m$, given a dataset $D = \{(\mathbf{x}_k, y_k)\}_{k=1}^N$. In this approach, classical data are embedded into an n-qubit Hilbert space through a quantum feature map $|\Phi(\mathbf{x})\rangle = U(\mathbf{x})\,|0\rangle^{\otimes n}$, where $U(\mathbf{x})$ is a parameterized encoding circuit. The similarity between two encoded inputs is

quantified by the fidelity kernel

$$K(\mathbf{x}, \mathbf{z}) = \left|\langle \Phi(\mathbf{x}) \mid \Phi(\mathbf{z}) \rangle\right|^2, \tag{1}$$

which can be estimated on a quantum device via the transition probability of the corresponding circuit [5], while the regression model itself is optimized classically.

To analyze the geometry of this mapping, we consider the corresponding density operators $\rho(\mathbf{x}) = |\Phi(\mathbf{x})\rangle\langle\Phi(\mathbf{x})|$. While the underlying n-qubit Hilbert space has complex dimension 2^n, the space of Hermitian operators acting on it constitutes a real vector space. This operator space is naturally spanned by the set of n-qubit Pauli operators $\mathcal{P}_n = \{I, X, Y, Z\}^{\otimes n}$ [20]. Consequently, any pure-state density matrix can be expanded in this basis as $\rho(\mathbf{x}) = \frac{1}{2^n} \sum_{i=1}^{d} a_i(\mathbf{x})\, \sigma^i$, where $\sigma^i \in \mathcal{P}_n$ and $a_i(\mathbf{x}) = \mathrm{tr}[\rho(\mathbf{x})\, \sigma^i]$ is the expectation value of the i-th Pauli operator. Using the orthogonality relation $\mathrm{tr}(\sigma^i \sigma^j) = 2^n \delta_{ij}$, the quantum kernel can be rewritten as

$$K(\mathbf{x}, \mathbf{z}) = \frac{1}{2^n} \sum_{i=1}^{d} a_i(\mathbf{x})\, a_i(\mathbf{z}), \tag{2}$$

where d denotes the total dimension of the Pauli feature space, typically $d = 4^n$ when the full Pauli basis is considered. This reformulation reveals that the quantum kernel is equivalent to a standard linear kernel in a real feature space of dimension 4^n, providing a direct geometric bridge between quantum and classical descriptions.

2.1 Axis-Aligned Least-Squares Upper Bound

Building upon this representation, we now formalize the least-squares regression model. Each encoded input $\mathbf{x}_k$ is mapped to a feature vector $\mathbf{a}(\mathbf{x}_k) = [a_1(\mathbf{x}_k), \ldots, a_d(\mathbf{x}_k)]$. Given a set of target labels $\{y_k\}_{k=1}^N$, the quality of any predictor $f : \mathcal{X} \to \mathbb{R}$ is measured by the empirical MSE, following the standard framework of empirical risk minimization [2]:

$$\mathrm{MSE}(f) = \frac{1}{N} \sum_{k=1}^{N} \big(y_k - f(\mathbf{x}_k)\big)^2. \tag{3}$$

We define the family of full affine linear regressors in the Pauli feature space as $\mathcal{F} = \{f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{a}(\mathbf{x}) \rangle + b \mid \mathbf{w} \in \mathbb{R}^d,\ b \in \mathbb{R}\}$. Ideally, we seek the predictor in $\mathcal{F}$ that minimizes Eq. (3), achieving the optimal training error $\mathrm{MSE}^* = \min_{f \in \mathcal{F}} \mathrm{MSE}(f)$. However, characterizing this minimum generally requires full access to the exponentially large feature space. As a tractable alternative, we consider the subclass of *axis-restricted* predictors, which depend on a single Pauli feature $i \in \{1, \ldots, d\}$:

$$\mathcal{F}^{(i)}_{\mathrm{axis}} = \Big\{ f_i(\mathbf{x}) = w\, a_i(\mathbf{x}) + b \;\Big|\; w, b \in \mathbb{R} \Big\}. \tag{4}$$

We denote the minimum error achievable by exploiting only the i-th axis as $\mathrm{MSE}_i = \min_{f \in \mathcal{F}_{\mathrm{axis}}^{(i)}} \mathrm{MSE}(f)$, and the best overall single-axis performance as $\mathrm{MSE}_{\mathrm{axis}} = \min_i \mathrm{MSE}_i$.

The nested structure of these hypothesis classes leads directly to a certified performance guarantee.

Theorem 1 (Pauli-axis upper bound) *For any dataset D encoded by a fixed quantum feature map, the optimal affine linear regression error is upper bounded by the best axis-aligned error:*

$$\mathrm{MSE}^* \leq \mathrm{MSE}_{\mathrm{axis}}.$$

Proof Since $\mathcal{F}_{\mathrm{axis}}^{(i)} \subset \mathcal{F}$ for every coordinate i, the minimum over the larger set $\mathcal{F}$ cannot exceed the minimum over the subset $\mathcal{F}_{\mathrm{axis}}^{(i)}$. Therefore, $\mathrm{MSE}^* \leq \mathrm{MSE}_i$ for all i. Taking the minimum over all d coordinates yields $\mathrm{MSE}^* \leq \min_i \mathrm{MSE}_i = \mathrm{MSE}_{\mathrm{axis}}$. □

This result extends naturally to more complex hypothesis classes $\mathcal{H} \supseteq \mathcal{F}$, such as those used in kernel ridge regression or support vector regression (SVR) on the induced feature space. Since these methods optimize over a space that includes all linear functions, their empirical error satisfies $\inf_{h \in \mathcal{H}} \mathrm{MSE}(h) \leq \mathrm{MSE}^*$. Consequently, $\mathrm{MSE}_{\mathrm{axis}}$ serves as a conservative, training-free proxy: if a simple axis-aligned model achieves low error, the more expressive kernel model is mathematically guaranteed to perform at least as well. However, to utilize this bound as a diagnostic tool, we require an explicit, computable expression for $\mathrm{MSE}_{\mathrm{axis}}$ in terms of observable data statistics.

2.2 Analytical Derivation and Complexity

We now derive an explicit analytical expression for the bound in Theorem 1. For any fixed Pauli coordinate i, the optimal predictor $f \in \mathcal{F}_{\mathrm{axis}}^{(i)}$ corresponds to the univariate ordinary least squares (OLS) solution. Its parameters depend solely on the empirical moments of the data. Let us define the sample means, variances, and covariances as:

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k, \qquad \bar{a}_i = \frac{1}{N}\sum_{k=1}^{N} a_i(\mathbf{x}_k),$$
$$\mathrm{Var}(a_i) = \frac{1}{N}\sum_{k=1}^{N} \big(a_i(\mathbf{x}_k) - \bar{a}_i\big)^2, \quad \mathrm{Cov}(a_i, y) = \frac{1}{N}\sum_{k=1}^{N} \big(a_i(\mathbf{x}_k) - \bar{a}_i\big)\big(y_k - \bar{y}\big). \tag{5}$$

If $\mathrm{Var}(a_i) > 0$, the optimal coefficients $w_i^\star$ and $b_i^\star$ are given by the standard OLS result:

$$w_i^\star = \frac{\mathrm{Cov}(a_i, y)}{\mathrm{Var}(a_i)}, \qquad b_i^\star = \bar{y} - w_i^\star \bar{a}_i. \tag{6}$$

Substituting these into the error function yields the residual variance:

$$\mathrm{MSE}_i = \mathrm{Var}(y) - \frac{\mathrm{Cov}(a_i, y)^2}{\mathrm{Var}(a_i)} = \mathrm{Var}(y)\big(1 - \rho_{y,a_i}^2\big), \tag{7}$$

where ρ_{y,a_i} is the Pearson correlation coefficient between the target y and the feature a_i. This equality follows from the fundamental property of simple linear regression, where the coefficient of determination R^2 is exactly the square of the correlation coefficient [21]. In the degenerate case where $\mathrm{Var}(a_i) = 0$, the predictor reduces to the constant mean $f_i(\mathbf{x}) = \bar{y}$, yielding $\mathrm{MSE}_i = \mathrm{Var}(y)$.

By substituting this result back into Theorem 1, we obtain a closed-form expression for the global upper bound:

$$\mathrm{MSE}_{\mathrm{axis}} = \min_{1 \le i \le d} \mathrm{MSE}_i = \mathrm{Var}(y) \left(1 - \max_{1 \le i \le d} \rho_{y,a_i}^2\right). \tag{8}$$

Eq. (8) offers a powerful geometric interpretation: the certified upper bound is determined solely by the single Pauli axis that exhibits the strongest marginal correlation with the target variable.

Regarding computational complexity, evaluating Eq. (8) requires iterating over all feature dimensions. Ideally, this consumes $O(Nd)$ time and $O(d)$ memory. However, two practical considerations arise. First, when Pauli expectations are estimated from finite quantum shots, the empirical moments inherit sampling noise. In statistical terms, this introduces measurement error in the regressors, leading to *attenuation bias* [22]: the noise artificially inflates $\mathrm{Var}(a_i)$ and systematically shrinks $|\rho_{y,a_i}|$. As a result, the computed $\mathrm{MSE}_{\mathrm{axis}}$ becomes conservative (larger than the ideal noiseless value), but preserves the validity of the upper bound.

Second, and more critically, the dimension d corresponds to the full operator basis size, $d = 4^n$. While the calculation is linear in d, the exponential growth of the basis with the number of qubits renders the exact evaluation of $\max_i \rho_{y,a_i}^2$ intractable for multi-qubit systems. This scaling bottleneck motivates the introduction of the Monte Carlo estimation framework presented in the next section.

3 Monte Carlo Estimation Framework

To overcome the exponential scaling of the Pauli basis ($d = 4^n$) identified in Section 2, we introduce a probabilistic estimation method. A similar Monte Carlo strategy has been recently proposed to estimate the minimum accuracy of quantum classifiers [23]. In this work, we extend and formalize this framework for the continuous regression domain, deriving specific concentration bounds for the MSE.

Instead of scanning all d coordinates, we uniformly sample a random subset of axes $T \subset \{1, \ldots, d\}$ of cardinality t, where $t \ll d$. We define the Monte Carlo estimator as the minimum error found within that subset:

$$\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T) := \min_{i \in T} \mathrm{MSE}_i = \mathrm{Var}(y) \left(1 - \max_{i \in T} \rho_{y,a_i}^2\right). \tag{9}$$

This estimator selects the best axis from the subset T, providing an efficiently computable surrogate for $\mathrm{MSE}_{\mathrm{axis}}$. Crucially, despite being an approximation, this estimator preserves the rigorous performance guarantee of the exact metric, as established by the following theorem.

Theorem 2 (Monte Carlo bound) *For any non-empty subset $T \subseteq \{1, \ldots, d\}$, the optimal affine linear error MSE^* is upper bounded by the Monte Carlo estimator*

$$\mathrm{MSE}^* \leq \mathrm{MSE}_{\mathrm{axis}} \leq \widehat{\mathrm{MSE}}_{\mathrm{axis}}(T). \tag{10}$$

Proof By definition,

$$\mathrm{MSE}_{\mathrm{axis}} = \min_{1 \leq j \leq d} \mathrm{MSE}_j, \qquad \widehat{\mathrm{MSE}}_{\mathrm{axis}}(T) = \min_{j \in T} \mathrm{MSE}_j.$$

Since $T \subseteq \{1, \ldots, d\}$ and $T \neq \varnothing$, monotonicity of the minimum implies

$$\mathrm{MSE}_{\mathrm{axis}} \leq \widehat{\mathrm{MSE}}_{\mathrm{axis}}(T).$$

Combining this inequality with Theorem 1, which states that $\mathrm{MSE}^* \leq \mathrm{MSE}_{\mathrm{axis}}$, yields

$$\mathrm{MSE}^* \leq \mathrm{MSE}_{\mathrm{axis}} \leq \widehat{\mathrm{MSE}}_{\mathrm{axis}}(T).$$

□

The inequalities in (10) confirm that the Monte Carlo estimator serves as a certified upper bound on the optimal error MSE^*, requiring no model training. Furthermore, this bound is systematically tightenable: the map $T \mapsto \widehat{\mathrm{MSE}}_{\mathrm{axis}}(T)$ is non-increasing with respect to set inclusion. Specifically, for nested subsets $T_t \subset T_{t+1}$, the estimator satisfies $\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T_{t+1}) \leq \widehat{\mathrm{MSE}}_{\mathrm{axis}}(T_t)$, converging to the exact value $\mathrm{MSE}_{\mathrm{axis}}$ as $t \to d$.

3.1 Statistical Guarantees and Threshold Selection

While Theorem 2 establishes that the Monte Carlo estimator $\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T)$ is a valid upper bound on the optimal affine regression error, it does not quantify the probability that this bound is sufficiently tight in practice. In regression tasks, tightness must be interpreted relative to the intrinsic scale of the target variable. For this reason, we introduce a target threshold τ defined in terms of a desired coefficient of determination $R^2_{\mathrm{target}} \in (0, 1]$:

$$\tau = \mathrm{Var}(y)\big(1 - R^2_{\mathrm{target}}\big). \tag{11}$$

Equivalently, we may write $\tau = \tau_{\mathrm{ratio}} \cdot \mathrm{Var}(y)$, where $\tau_{\mathrm{ratio}} := 1 - R^2_{\mathrm{target}}$.

Any axis i satisfying $\mathrm{MSE}_i \leq \tau$ therefore explains at least a fraction R^2_{target} of the sample variance. To characterize how frequently such axes occur in the Pauli feature space, we define the set of *good axes*

$$\mathcal{A}_\tau = \{i \in \{1, \ldots, d\} : \mathrm{MSE}_i \leq \tau\}, \tag{12}$$

and introduce the empirical cumulative distribution function over the finite axis set,

$$F(\tau) = \frac{1}{d}\,|\mathcal{A}_\tau|. \tag{13}$$

We denote by $p := F(\tau)$ the fraction of axes achieving the target performance level.

The following result quantifies the probability that a random subset of axes captures at least one good axis.

Theorem 3 *Fix a threshold τ and let $p = F(\tau)$. If $T \subset \{1, \dots, d\}$ is sampled uniformly without replacement with $|T| = t$, then*

$$\mathbb{P}\Big(\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T) \le \tau\Big) = 1 - \frac{\binom{d-\lfloor pd \rfloor}{t}}{\binom{d}{t}} \;\ge\; 1 - (1-p)^t. \tag{14}$$

Proof Let $\mathcal{A}_\tau$ denote the set of good axes with cardinality $m = |\mathcal{A}_\tau| = \lfloor pd \rfloor$. The event $\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T) \le \tau$ occurs if and only if $T \cap \mathcal{A}_\tau \neq \emptyset$. Under uniform sampling without replacement, the probability that T avoids $\mathcal{A}_\tau$ is given by the hypergeometric term $\binom{d-m}{t}/\binom{d}{t}$. The inequality follows from the bound $\binom{d-m}{t}/\binom{d}{t} \le (1 - m/d)^t = (1-p)^t$. □

An immediate consequence of Theorem 3 is a sufficient condition on the sample size required to achieve a desired confidence level.

Corollary 4 *Fix $\delta \in (0,1)$ and suppose that at least a fraction $p \in (0,1)$ of the axes satisfies $\mathrm{MSE}_i \le \tau$. If T is sampled uniformly without replacement with $|T| = t$, then the condition*

$$t \;\ge\; \frac{\log(1/\delta)}{\log\!\big(1/(1-p)\big)} \tag{15}$$

ensures

$$\mathbb{P}\Big(\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T) \le \tau\Big) \ge 1 - \delta.$$

In particular, since $-\log(1-p) \ge p$ for all $p \in (0,1)$, the simpler sufficient condition

$$t \;\ge\; \frac{1}{p}\,\log\frac{1}{\delta} \tag{16}$$

also guarantees the same confidence level.

Proof By Theorem 3, $\mathbb{P}(\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T) \le \tau) \ge 1 - (1-p)^t$. Requiring the right-hand side to be at least $1 - \delta$ yields $(1-p)^t \le \delta$, which is equivalent to (15). The bound (16) follows from $-\log(1-p) \ge p$. □

3.2 Adaptive Sample-Size Calibration

The sample-size bounds in Corollary 4 depend on the unknown fraction $p = F(\tau)$ of axes satisfying the threshold condition. Since computing p via an exhaustive scan of the $d = 4^n$ Pauli axes is intractable at scale, we employ an adaptive Monte Carlo procedure that estimates this quantity online while preserving rigorous probabilistic guarantees.

Fixing a threshold τ, each sampled axis i is associated with the indicator variable

$$X_i := \mathbf{1}\{\mathrm{MSE}_i \le \tau\}, \tag{17}$$

which records whether the axis meets the target performance. Although sampling is performed without replacement from a finite population, we employ one-sided Hoeffding bounds, which remain conservative under this regime.

After sampling a subset T of t distinct axes, define

$$s := \sum_{i \in T} X_i, \qquad \hat{p} := \frac{s}{t}, \tag{18}$$

as the number of successful axes and the corresponding empirical success rate. For a confidence parameter $\alpha \in (0, 1)$, a lower confidence bound on p is given by

$$p_L := \max\left\{0,\ \hat{p} - \sqrt{\frac{1}{2t}\log\frac{1}{\alpha}}\right\}, \tag{19}$$

which holds with probability at least $1 - \alpha$.

Substituting p_L into the exact coverage condition of Theorem 3 yields a data-driven estimate of the required sample size,

$$t_{\mathrm{req}} := \frac{\log(1/\delta)}{\log\big(1/(1 - p_L)\big)}, \tag{20}$$

with the convention $t_{\mathrm{req}} = \infty$ when $p_L = 0$. Sampling proceeds iteratively until the stopping condition $t \ge t_{\mathrm{req}}$ is met, at which point the estimator satisfies $\mathbb{P}(\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T) \le \tau) \ge 1 - \delta$.

To prevent unbounded execution in regimes where high-quality axes are extremely sparse, we impose a maximum sampling budget $t_{\max}$. In addition, a futility stopping criterion is employed: if no satisfactory axes are observed ($\hat{p} = 0$) and the confidence width falls below a minimal tolerance, the procedure terminates early. In such cases, the returned value $\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T)$ remains a valid certified upper bound on the optimal regression error by Theorem 2, albeit without the threshold guarantee. This adaptive strategy, summarized in Algorithm 1, translates the theoretical coverage bounds into a practical and robust stopping rule, allocating computational effort in proportion to the empirical density of high-performing axes.

Algorithm 1 Adaptive Monte Carlo calibration of the axis-aligned bound

Require: τ_{ratio}, δ_{total}, t_0, b, $t_{\max}$, $\epsilon_{\min}$
Ensure: Sampled set T, estimator $\widehat{\text{MSE}}_{\text{axis}}(T)$

$\tau \leftarrow \tau_{\text{ratio}}\text{Var}(y), \quad \alpha = \delta = \delta_{\text{total}}/2$
Sample initial set T ($|T| = t_0$); $\quad t \leftarrow t_0$, $s \leftarrow \sum_{i \in T} \mathbf{1}\{\text{MSE}_i \leq \tau\}$
while $t < t_{\max}$ **do**
  $\hat{p} \leftarrow s/t, \quad \epsilon \leftarrow \sqrt{\frac{1}{2t}\log\frac{1}{\alpha}}, \quad p_L \leftarrow \max\{0, \hat{p} - \epsilon\}$
  $t_{\text{req}} \leftarrow \log(1/\delta)/\log(1/(1 - p_L))$
  if $t \geq t_{\text{req}}$ **or** $(\hat{p} = 0 \wedge \epsilon < \epsilon_{\min})$ **then break**
  end if
  Update T with b new axes; $\quad s \leftarrow s + \Delta s, \quad t \leftarrow t + b$
end while
return T, $\min_{i \in T} \text{MSE}_i$

Algorithm 1 formalizes the adaptive Monte Carlo procedure used throughout the experimental section, providing a certified stopping rule that avoids exhaustive scans of the 4^n axis-aligned hypothesis space.

4 Experimental Setup

To validate the theoretical framework and the adaptive Monte Carlo procedure, we designed a comprehensive experimental suite varying structural complexity, noise levels, and feature map architectures.

4.1 Datasets and Benchmarking Strategy

To rigorously validate the behavior of the adaptive estimator, we employ a hybrid benchmarking strategy. First, we construct two synthetic landscapes with diametrically opposed geometric properties, one dense and correlated, the other sparse and high-dimensional, to serve as controlled stress tests. Since the ground truth structure of these tasks is known by design, they allow us to verify if the estimator's convergence behavior aligns with theoretical expectations. Finally, we challenge the method on a real-world dataset to assess its robustness under unstructured noise.

The three representative regression tasks are summarized in Table 1: (i) `synthetic_corr_gauss`, representing a moderate regime with correlated features; (ii) `synthetic_sparse`, a hard regime with high-dimensional sparse dependencies; and (iii) `real_california`, a subset of the California Housing dataset reduced to $n = 4$ principal components, serving as a real-world benchmark with natural noise.

4.2 Quantum Feature Map and Configuration Space

The quantum feature space is generated by the Pauli Feature Map ansatz [5], which maps classical data $\mathbf{x} \in \mathbb{R}^n$ into an n-qubit state via interleaved Hadamard layers and

Table 1 Profile of datasets used in the numerical experiments.

Property	Details
Dataset	`synthetic_corr_gauss`
Regime	**Moderate (Dense):** Information is spread across correlated axes; high density of near-optimal solutions.
Target Function	$y = \exp(-\frac{1}{2}\mathbf{x}^\mathsf{T}\Sigma^{-1}\mathbf{x})$, with $\Sigma_{ij} = 0.3 + 0.7\delta_{ij}$.
Dataset	`synthetic_sparse`
Regime	**Hard (Sparse):** Information is concentrated in a hidden subset; low density of informative axes (needle-in-haystack).
Target Function	$y = \sum_{i\in\mathcal{S}} \sin(x_i) + 0.1\sum_{j\notin\mathcal{S}} x_j$, where $\mathcal{S}$ is the active subset.
Dataset	`real_california`
Regime	**Real-world:** Natural data distribution with unstructured noise and complex dependencies.
Description	California Housing (PCA reduced to $n = 4$, retaining 95% variance).

data-dependent entangling gates:

$$\mathcal{U}_{\Phi(\mathbf{x})} = \left(U_{\Phi(\mathbf{x})}\, H^{\otimes n}\right)^r, \qquad U_{\Phi(\mathbf{x})} = \exp\left(i \sum_{S\in\mathcal{I}} \phi_S(\mathbf{x}) \prod_{k\in S} P_k\right). \tag{21}$$

Here, $P_k \in \{X, Y, Z\}$ are Pauli operators, $\mathcal{I}$ defines the connectivity graph, and $\phi_S(\mathbf{x})$ is a classical encoding function.

To ensure robustness and architectural diversity, our study does not rely on a single model. Instead, we conduct a comprehensive evaluation across a total of 324 experimental configurations, comprising 108 distinct feature maps applied to each dataset. We systematically explore this design space by permuting the three key components defined in Table 2: (1) *classical preprocessing*, to handle data scaling; (2) *Pauli sequences*, determining the basis of quantum correlations; and (3) *data-mapping rules*, controlling the nonlinearity of the phase encoding. These core components are further combined with varying entanglement patterns and circuit depths ($r \in \{1, 2\}$) to cover a broad spectrum of model complexity.

5 Numerical Performance and Benchmarking

This section assesses the predictive power of the certified bound by comparing it against the performance of fully trained regression models. Our primary goal is to demonstrate that the training-free estimator $\widehat{\mathrm{MSE}}_{\mathrm{axis}}(T)$ effectively identifies high-quality feature maps without requiring expensive variational optimization.

For $n = 4$ qubits ($d = 256$), we first performed a training-free scan across the grid of 108 configurations per dataset. The adaptive estimator was configured with a confidence level $\delta_{\mathrm{total}} = 0.05$ and a relative threshold $\tau_{\mathrm{ratio}} = 0.95$. Sampling was initialized with a pilot size $t_0 = 50$ and updated in batches of $b = 20$, up to a maximum budget $t_{\max} = d$.

Table 2 Configuration space definition. The Pauli sequences include `all_pairs` which comprises all two-body combinations (`XY`, `YZ`, `ZZ`, etc.).

Stage	Option and definition	Qualitative effect
Classical preprocessing	`id`: $f(x) = x$	Unbounded inputs (raw data)
	`tanh`: $f(x) = \tanh(x)$	Bounded to $[-1, 1]$, outlier-robust
	`rbf-s1`: $f(x) = e^{-x^2/2}$	Local similarity emphasis (Gaussian)
Pauli sequence	`Z+ZZ` ([Z, ZZ])	Computational basis correlations
	`Y+YY` ([Y, YY])	Off-diagonal interference (complex)
	`all_pairs` ([XY, YZ, ZZ...])	Richer entanglement structure
Data-mapping ($\phi(x)$)	`prod`: $\prod_i x_i$	Purely nonlinear cross-terms
	`pi*prod`: $\pi \prod_i x_i$	Full 2π phase coverage
	`sum+prod`: $\sum_i x_i + \prod_i x_i$	Mixture of additive and multiplicative terms

Table 3 Anchor configurations: Comparison between the training-free Monte Carlo bound ($\widehat{\mathrm{MSE}}_{\mathrm{axis}}$), the exact single-axis optimum ($\mathrm{MSE}_{\mathrm{axis}}$), and fully trained models.

Dataset	Configuration	$\mathrm{MSE}_{\mathrm{axis}}$	$\widehat{\mathrm{MSE}}_{\mathrm{axis}}$	t_{used}	$\mathrm{MSE}_{\mathrm{ridge}}$	$\mathrm{MSE}_{\mathrm{svr}}$
`real_california`	`rbf-s1 | prod | Z+ZZ | lin | r=2`	0.6863	0.6863	110	0.4123	0.5105
`syn_corr_gauss`	`rbf-s1 | prod | Y+YY | lin | r=1`	0.3195	0.7053	50	0.1165	0.0101
`syn_sparse`	`tanh | prod | Z+ZZ | full | r=1`	0.2684	0.2684	256*	0.0000	0.0094

*Stopped by budget limit.

For each dataset, we identified the single best-performing architecture, referred to as the "anchor" configuration, based solely on the adaptive estimator. To validate this selection, we then trained two regression models specifically for these anchors:

1. **Quantum Ridge Regression ($\mathrm{MSE}_{\mathrm{ridge}}$):** A linear regressor with ℓ_2 regularization ($\alpha = 10^{-3}$) trained directly on the d-dimensional feature vector $\Phi(\mathbf{x})$. This serves as an ideal proxy for a noise-free Quantum Support Vector Regressor (QSVR).
2. **Classical SVR ($\mathrm{MSE}_{\mathrm{svr}}$):** A standard Support Vector Regressor with an RBF kernel trained on the original classical data, serving as a baseline.

5.1 Results and Discussion

Table 3 summarizes the results for the best "anchor" configuration identified for each dataset (see Table 1 for dataset profiles). The results reveal three key insights regarding the alignment between the certified bound, the dataset geometry, and the actual training potential:

1. Real-World Complexity and Linear Combinations.

For `real_california`, the estimator converged efficiently ($t = 110 < d/2$) and matched the exhaustive bound. However, a significant gap remains between the best single-axis feature ($\text{MSE}_{\text{axis}} \approx 0.68$) and the fully trained Quantum Ridge model ($\text{MSE}_{\text{ridge}} \approx 0.41$). This behavior is consistent with the "Real-world" regime description: while the adaptive procedure correctly identified the best individual Pauli features, the superior performance of the trained regressor confirms that solving realistic tasks requires the linear combination of multiple features (superposition) rather than a single optimal axis. Notably, the Quantum Ridge outperformed the Classical SVR (0.51), suggesting the Pauli feature map captures non-trivial correlations in the housing data better than the standard RBF kernel.

2. Feature Map Alignment in the Sparse Regime.

In the `syn_sparse` case, the Ridge regressor achieved perfect reconstruction (MSE $\approx$ 0.0000), theoretically validating that the chosen feature map (`Z+ZZ`) perfectly spans the generating function of the dataset. The adaptive estimator provided a crucial diagnostic here: although it exhausted the budget ($t = 256$), confirming the "Hard/Sparse" regime where good axes are rare needles in a haystack, it successfully recovered the optimal axis (Gap = 0.0). This proves the method's robustness: even when certification is statistically difficult due to sparsity, the search procedure is still effective at locating high-quality features for optimization.

3. The "Dense Trap" in Correlated Data.

The `syn_corr_gauss` dataset illustrates the nuance of the "Moderate/Dense" regime. The estimator stopped early ($t = 50$) with a loose bound. This is not a failure, but a consequence of the high density of "reasonably good" axes in correlated landscapes. The algorithm quickly satisfied the threshold condition τ and stopped, as designed, before stumbling upon the much rarer global optimum. Furthermore, the Classical SVR (0.0101) dominated this task, which is expected since the target is a smooth Gaussian function, naturally aligned with the classical RBF kernel but harder to approximate with discrete Pauli strings (Ridge ≈ 0.11).

In summary, the Monte Carlo estimator acts as a cost-effective probe: it correctly flagged the sparsity of `syn_sparse` (via budget exhaustion) and the density of `syn_corr_gauss` (via early stopping), while providing a tight predictive bound for the real-world scenario.

6 Conclusion

In this work, we introduced a *certified, training-free* upper bound on the optimal training error for quantum kernel regression. By extending the axis-aligned framework of Suzuki *et al.* from classification to the regression setting, we derived a closed-form metric that links the geometry of the Pauli feature space directly to the mean squared error. To address the exponential growth of the feature dimension, we proposed an adaptive Monte Carlo estimator equipped with rigorous statistical guarantees based on Hoeffding concentration inequalities.

Extensive numerical experiments on both synthetic and real-world datasets confirmed the efficiency and predictive power of the method. The estimator consistently converged to the true exhaustive bound using a sampling budget substantially smaller than the full basis dimension, across both sparse and dense landscapes. Moreover, comparisons against fully trained Ridge regressors demonstrated that the certified bound is not merely a loose theoretical limit, but a high-fidelity predictor of the expressive capacity of a given feature map, enabling reliable discrimination between high- and low-performing architectures without the cost of explicit training or regression fitting.

While the certification procedure is inherently threshold-dependent, this dependence should be viewed as a feature rather than a limitation. In our experiments, fixed hyperparameters were adopted to ensure fair comparisons across datasets; however, parameters such as the threshold ratio τ_{ratio} and the confidence level δ_{total} offer principled control over the trade-off between selectivity and sampling cost. This flexibility allows practitioners to adapt the certification process to dataset-specific characteristics and available computational resources.

Overall, the proposed framework establishes a practical diagnostic tool for *Quantum Model Selection* in the NISQ era. By filtering out poor feature-map configurations prior to training, it enables computational effort to be focused exclusively on the most promising quantum architectures. A natural direction for future work is to exploit the subset of informative axes identified by the adaptive procedure as a compressed feature space, potentially guiding or constraining variational optimization while preserving low computational overhead.

References

[1] Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th edn. Pearson Series in Artificial Intelligence. Pearson, Boston, MA (2020)

[2] Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York, NY (2006)

[3] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S.: Quantum machine learning. Nature **549**(7671), 195–202 (2017)

[4] Grant, E., Benedetti, M., Cao, S., Hallam, A., Lockhart, J., Stojevic, V., Green, A.G., Severini, S.: Hierarchical quantum classifiers. npj Quantum Information **4**(1), 65 (2018)

[5] Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., Gambetta, J.M.: Supervised learning with quantum-enhanced feature spaces. Nature **567**(7747), 209–212 (2019)

[6] Schuld, M., Killoran, N.: Quantum machine learning in feature Hilbert spaces. Physical Review Letters **122**(4), 040504 (2019)

[7] Mitarai, K., Negoro, M., Kitagawa, M., Fujii, K.: Quantum circuit learning.

Physical Review A **98**(3), 032309 (2018) https://doi.org/10.1103/PhysRevA.98.032309

[8] Huang, H.-Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H., McClean, J.R.: Power of data in quantum machine learning. Nature Communications **12**(1), 2631 (2021)

[9] Li, W., Lu, Z.: Quantum kernel method for high-dimensional data analysis. IEEE Transactions on Quantum Engineering **3**, 1–11 (2022)

[10] Tacchino, F., Macchiavello, C., Gerace, D., Bajoni, D.: Quantum implementation of an artificial neural network. npj Quantum Information **5**(1), 26 (2019)

[11] Gentinetta, G., Thomsen, A., Sutter, D., Woerner, S.: The complexity of quantum support vector machines. Quantum **8**, 1225 (2024) https://doi.org/10.22331/q-2024-01-11-1225

[12] McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. Nature Communications **9**(1), 4812 (2018) https://doi.org/10.1038/s41467-018-07090-4

[13] Cerezo, M., Sone, A., Volkoff, T., Cincio, L., Coles, P.J.: Cost function dependent barren plateaus in shallow parametrized quantum circuits. Nature Communications **12**(1), 1791 (2021) https://doi.org/10.1038/s41467-021-21728-w

[14] Larocca, M., Cerezo, M., Sharma, K., Sone, A., Coles, P.J., Holmes, Z.: Barren plateaus in variational quantum computing. Reviews of Modern Physics (2024). Preprint at arXiv:2205.05756

[15] Sim, S., Johnson, P.D., Aspuru-Guzik, A.: Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. Advanced Quantum Technologies **2**(12), 1900070 (2019) https://doi.org/10.1002/qute.201900070

[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

[17] Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing **14**(3), 199–222 (2004) https://doi.org/10.1023/B:STCO.0000035301.49549.88

[18] Suzuki, Y., Yano, H., Gao, Q., Uno, S., Tanaka, T., Akiyama, M., Yamamoto, N.: Analysis and synthesis of feature map for kernel-based quantum classifier. Quantum Machine Intelligence **4**(1), 1–15 (2022) https://doi.org/10.1007/s42484-022-00058-8

[19] Hoeffding, W.: Probability inequalities for sums of bounded random variables.

Journal of the American Statistical Association **58**(301), 13–30 (1963) https://doi.org/10.1080/01621459.1963.10500830

[20] Nielsen, M.A., Chuang, I.L.: Quantum Computation and Quantum Information: 10th Anniversary Edition. Cambridge University Press, Cambridge (2010)

[21] Kutner, M.H.: Applied Linear Statistical Models. McGrwa-Hill international edition. McGraw-Hill Irwin, New York (2005)

[22] Fuller, W.A.: Measurement Error Models. Wiley Series in Probability and Statistics. John Wiley & Sons, New York (1987). https://doi.org/10.1002/9780470316665

[23] Gonçalves, D.N., Fernandes, T.D., Cordeiro, A.M.M., Lugao, P.H.G., Dias, J.T.: Certified Lower Bounds and Efficient Estimation of Minimum Accuracy in Quantum Kernel Methods (2025). https://arxiv.org/abs/2512.20588