SPECTRAL ANATOMY OF QUANTUM GAUSSIAN PROCESS KERNELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Two recent results have reshaped quantum Gaussian processes (QGPs). On the one hand, Lowe et al. (2025) rule out the exponential speedups claimed by HHL-based QGP regression in the typical regime; on the other, an independent line of work shows that highly expressive quantum kernels suffer posterior pathologies that break Bayesian optimization. We show that these seemingly unrelated phenomena are governed by the same quantity: the normalized spectral entropy $S(K)/\log n$ of the kernel Gram matrix. We prove a Cauchy–Schwarz bound on the best rank-k spectral truncation error, a finite-sample variance-contraction bound, and a characterization of the *target-dependent* optimal entropy via the intrinsic dimension of the target in the kernel eigenbasis. Empirically, the diagnostic is kernel-agnostic: hardware-efficient, matchgate, IQP *and* RBF/Matérn/RFF/deep-kernel families all collapse onto identical $S/\log n$ curves on dequantization, ECE, and variance-contraction panels. The NLL sweet spot lives at high entropy for smooth targets and at low entropy for band-limited quantum-data targets. The diagnostic transfers from simulator to IBM Heron hardware with median absolute error 3.2% and mean 5.2% in $S/\log n$ across 24 configurations at $n_q = 4$, with matchgate and IQP within 5% mean and a single HE configuration returning a 30% outlier that drops to 0.5% on rerun (attributed to calibration drift); the same diagnostic transfers to a second Heron backend (mean error 2.7%) and to a $n_q = 6$ scale-up on the original backend (mean error 1.7%). No error mitigation is applied throughout.

1 INTRODUCTION

Quantum Gaussian processes (QGPs) promise two things that classical GPs cannot easily deliver: kernels that encode physics-motivated symmetries (Jäger et al., 2026; Rapp & Roth, 2024), and—in principle—asymptotic runtime speedups for Bayesian inference via quantum linear solvers (Zhao et al., 2019a;b; Chen et al., 2022). The past eighteen months have however brought a sharp reframing.

First, Lowe et al. (2025) proved that the kernel matrices arising in many QGP regression settings—specifically, those for which HHL-based exponential speedups had previously been claimed—have condition number scaling $\kappa(K) = \Omega(n)$, so that the HHL pipeline delivers at most polynomial advantage in those regimes. Second, an independent line of work on quantum-kernel concentration (Thanasilp et al., 2024; Kübler et al., 2021) shows that highly expressive quantum feature maps produce kernel matrices that approach either the constant matrix $c\mathbf{1}\mathbf{1}^\top$ or the identity I_n, causing the GP posterior to either over-fit confidently to garbage or collapse to the prior. Both pathologies destroy downstream uses of the posterior such as Bayesian optimization (Dai et al., 2023; Rapp & Roth, 2024).

Each finding has been treated as a separate cautionary tale. We argue they are the same tale, told from different angles. *The eigenspectrum of the kernel Gram matrix* controls both: how well a classical random-feature scheme can approximate the kernel, and how informative the GP posterior remains. A small number of spectral statistics, computable from training data alone, predict both phenomena across ansatz families. They identify a Goldilocks region of *useful hardness*: kernels that are not classically trivial, but also not so concentrated that the GP posterior dies. The location of this region is not fixed; it depends on the target function.

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

Contributions. We make three concrete claims, all backed by quantitative experiments on three ansatz families and a real IBM Heron device.

(C1) A label-free spectral coordinate. We adopt the normalized spectral entropy $S(K)/\log n$ as a single *scale-invariant, label-free* design coordinate on which dequantization difficulty, posterior calibration, and variance contraction can be read off jointly (Sections 3–4). Structurally diverse quantum families (hardware-efficient, matchgate, IQP) and classical baselines (RBF/Matérn/RFF/deep) populate *the same narrow band on this coordinate* once (L, s) is swept, which we use as the empirical basis for treating $S/\log n$ as the natural axis for downstream design.

(C2) The target-dependent half. The location of the predictive NLL minimum on the spectral axis depends on the target. Classically expressive targets place the minimum at high entropy ($S/\log n \approx 0.9$); structured quantum-data targets place it at low entropy ($S/\log n \approx 0.1$). The frontier is shaped by the interaction of kernel spectrum and target spectral content (Section 5).

(C3) NISQ-portable estimation. The spectral statistics defining the diagnostic can be estimated on current IBM Heron processors with errors compatible with simulator sampling variation: across 24 frontier configurations at $n_q = 4$ on `ibm_aachen`, the hardware-vs.-simulator deviation in $S/\log n$ has median 0.032 and mean 0.052, with matchgate and IQP within 0.093 worst-case; a single HE configuration returned 0.300 in one run and 0.005 on rerun, which five repeated reruns attribute to backend calibration drift (Section 6). A second Heron device (`ibm_marrakesh`) and a $n_q = 6$ scale-up (mean error 0.017) give consistent numbers. All hardware results use *no error mitigation*, so the reader reads off the raw cost of moving the diagnostic to a device.

Section 7 turns this into a practical recipe that, from an ansatz and unlabeled inputs alone, returns the operating regime.

2 Background and Related Work

Quantum GPs, speedups, and their retraction. The seminal QGP algorithms build posterior mean and variance estimates via quantum linear systems (Zhao et al., 2019a;b; Harrow et al., 2009), with later work relaxing the oracular-kernel assumption (Chen et al., 2022) or exploiting reduced rank (Farooq et al., 2024; Galvis-Florez et al., 2025). Following the dequantization program of Tang (2019), Lowe et al. (2025) prove that the kernel matrix in many QGP settings has $\kappa(K) = \Omega(n)$, implying at most polynomial advantage and relocating the question from "is there a speedup" to "what spectral structure makes a kernel non-trivially quantum." A parallel theoretical line shows that quantum neural networks converge to GPs in suitable limits (Garcıa-Martın et al., 2023; Girardi & De Palma, 2025; Melchor Hernandez et al., 2025), and Jäger et al. (2026) construct scalable, provable QGPs from matchgate symmetry.

Kernel concentration. Quantum kernels from deep random circuits suffer exponential concentration of measure (Thanasilp et al., 2024; Kübler et al., 2021), the kernel analogue of barren plateaus (McClean et al., 2018). Concentration is typically diagnosed by the variance of off-diagonal kernel entries; we show in Section 3 that this variance is a coarse proxy for the spectral statistics that more directly control downstream behavior. On hardware, QGP regression and quantum-kernel Bayesian optimization have been demonstrated at small scale (Otten et al., 2020; Rapp & Roth, 2024; Dai & Krems, 2022).

Gap. Existing QGP literature treats kernel-dequantizability and kernel-concentration as separate phenomena tied to different sub-fields (quantum complexity vs. trainability). No prior work analyzes them via a single spectral framework, and no prior work isolates the role of the target in selecting the useful spectral regime. The present paper supplies both; Appendix C gives the extended discussion.

3 THE SPECTRAL FRAMEWORK

3.1 SETTING AND SPECTRAL STATISTICS

Definition 1 (Quantum Gaussian process)**.** Let $U_\phi : \mathcal{X} \to \mathrm{U}(2^{n_q})$ be a parameterized unitary on n_q qubits with input domain $\mathcal{X} \subseteq \mathbb{R}^d$. The associated *quantum kernel* is

$$K_q(x, x') \;=\; \big|\langle 0^{n_q}|U_\phi(x')^\dagger U_\phi(x)|0^{n_q}\rangle\big|^2.$$

Given training pairs $(X, y) = (\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$ with $y_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, the GP regressor $\mathrm{GP}(0, K_q)$ has predictive mean and variance at x_*

$$\mu(x_*) = k_*^\top (K + \sigma^2 I)^{-1} y, \qquad v(x_*) = K_q(x_*, x_*) - k_*^\top (K + \sigma^2 I)^{-1} k_*,$$

where $K_{ij} = K_q(x_i, x_j)$ and $(k_*)_i = K_q(x_i, x_*)$.

Definition 2 (Spectral statistics)**.** For PSD $K \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ and normalized eigenvalues $\tilde{\lambda}_i := \lambda_i / \sum_j \lambda_j$, define

$$\begin{aligned} r_{\text{eff}}(K) \;&=\; \frac{\big(\sum_i \lambda_i\big)^2}{\sum_i \lambda_i^2} \;\in\; [1, n], \\ S(K) \;&=\; -\sum_i \tilde{\lambda}_i \log \tilde{\lambda}_i \;\in\; [0, \log n], \\ c(K) \;&=\; \mathrm{std}(K_{\text{off}}) \,/\, \mathrm{mean}|K_{\text{off}}|, \end{aligned}$$

namely the participation ratio, the normalized Shannon entropy, and the off-diagonal concentration proxy.

All three are computable from K alone. We adopt $s(K) := S(K)/\log n \in [0, 1]$ as the primary diagnostic; the next lemma shows that r_{eff} is essentially redundant.

Lemma 1 (Entropy controls effective rank)**.** *For any PSD K with eigenvalues as above,*

$$1 \;\leq\; r_{\text{eff}}(K) \;\leq\; e^{S(K)} \;\leq\; n^{s(K)} \;\leq\; n.$$

The first inequality is tight at constant collapse; the third is tight at Haar concentration; the bound $r_{\text{eff}} \leq e^S$ is the standard Rényi-2-vs.-Shannon entropy inequality.

3.2 TWO REFERENCE PATHOLOGIES

The two ends of the entropy axis are degenerate. If K is a perturbed constant matrix $c\mathbf{1}\mathbf{1}^\top + \delta E$ (and the test cross-covariance shares the same constant structure) then $r_{\text{eff}}(K) = 1 + O(\delta/c)$ and $S(K) = O((\delta/c)\log(c/\delta))$, and, for any fixed $\sigma^2 > 0$, the GP posterior degenerates to a *shrunk* empirical mean $\mu(x_*) = \frac{nc}{nc+\sigma^2}\bar{y} + O_\sigma(\delta/c)$ with an x_*-independent variance (Lemma 2; the shrinkage factor $nc/(nc + \sigma^2) < 1$ is a persistent regularization bias, not an artifact of δ, and σ^2 must be fixed away from 0—see the lemma for a counterexample to the exactly noiseless case). If instead $K = cI_n + \delta E$ with small test–train cross-covariance, then $r_{\text{eff}}(K) = n + O(n\delta/c)$, $S(K) = \log n + O(\delta/c)$, and the posterior collapses to the prior at every point outside the training set (Lemma 3). Both statements are proved in Appendix D. Both pathologies are *simultaneously* classically simulable—the first by the constant feature, the second by Gaussian random features approximating the identity—*and* useless for GP inference. The useful QGP regime must therefore live strictly between $s(K) = 0$ and $s(K) = 1$.

3.3 DEQUANTIZATION, CALIBRATION, AND CONTRACTION PROXIES

Definition 3 (Trace-truncation rank)**.** For $\varepsilon \in (0, 1)$, the dequantization proxy

$$\mathrm{dq}_\varepsilon(K) \;:=\; \min\Big\{k : \sum_{i=1}^k \tilde{\lambda}_i \geq 1 - \varepsilon\Big\} \,/\, n$$

is the fractional rank capturing $1 - \varepsilon$ of the trace. We use $\varepsilon = 0.01$.

Theorem 1 (Cauchy–Schwarz tail bound; high-rank regime)**.** *For any PSD K with sorted normalized eigenvalues $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ and any $k \in \{0, 1, \ldots, n\}$, the nuclear-norm tail satisfies*

$$\frac{\|K - K_k\|_*}{\|K\|_*} = \sum_{i>k} \tilde{\lambda}_i \leq \sqrt{\frac{n-k}{r_{\text{eff}}(K)}}.$$

In particular, if $r_{\text{eff}}(K) \geq \rho\, n$ for some $\rho \in (0, 1]$, then fixing $\varepsilon \in (0, 1)$ the rank-k truncation with $n - k \leq \rho\, n\, \varepsilon^2$ achieves nuclear-norm tail error $\leq \varepsilon$: keeping $k \geq n(1 - \rho\varepsilon^2)$ directions is sufficient *for ε-accurate compression in this high-rank regime. This is a sufficient, not a necessary, condition on k—see the discussion after the proof for how loose it can be.*

Proof sketch. $\sum_{i>k} \tilde{\lambda}_i \leq \sqrt{(n-k)\sum_{i>k} \tilde{\lambda}_i^2} \leq \sqrt{(n-k)\sum_i \tilde{\lambda}_i^2} = \sqrt{(n-k)/r_{\text{eff}}(K)}$ by Cauchy–Schwarz. Substituting $r_{\text{eff}} \geq \rho n$ gives the stated rank-deficiency bound. Full statement in Appendix H. □

What the bound does and does not say. Theorem 1 is a *sufficient* condition, not a matching lower bound: it says that keeping $k \geq n(1 - \rho\varepsilon^2)$ directions suffices for ε-accurate compression when $r_{\text{eff}} \geq \rho n$, but it does not say that fewer directions cannot also suffice. Indeed for $K = I$ (the Haar limit, $r_{\text{eff}} = n$, $\rho = 1$), the true relative tail is *exactly* $(n - k)/n$, so any $k \geq n(1 - \varepsilon)$ already achieves ε-accurate compression—an $O(\varepsilon)$ fraction discarded, quadratically more than the $O(\varepsilon^2)$ fraction the Cauchy–Schwarz bound guarantees. The bound is loose because Cauchy–Schwarz is tight only when the tail eigenvalues are already comparable in magnitude to the largest one it is bounding against; the qualitative story "high entropy is hard to compress" is not itself a consequence of this theorem and should be read as the empirical dq trend of Figures 4, 2, not as a proven necessary condition. Symmetrically, the theorem does *not* establish the reverse direction "low spectral entropy implies low-rank Nyström compressibility"—in fact, Cauchy–Schwarz on $\sum \tilde{\lambda}_i^2$ gives a *larger* upper bound on the tail as r_{eff} shrinks, so the bound is trivial in the low-r_{eff} regime too. A tight theoretical guarantee in either regime requires additional spectral-decay assumptions (e.g. polynomially decaying eigenvalues), which we leave for follow-up work; the empirically observed dq behavior is reported as an empirical characterization, not as a theorem.

Proposition 1 (Variance contraction; finite-sample bound)**.** *Let* $\mathrm{VC}(K) := \frac{1}{n_*} \sum_{j=1}^{n_*} v(x_*^{(j)})/K_q(x_*^{(j)}, x_*^{(j)})$ *be the average posterior variance contraction over a test set $\{x_*^{(j)}\}_{j=1}^{n_*}$. Then*

$$\mathrm{VC}(K) \geq 1 - \frac{1}{n_* \min_j K_q(x_*^{(j)}, x_*^{(j)})} \operatorname{tr}\Big(K_*^\top (K + \sigma^2 I)^{-1} K_*\Big).$$

This bound uses the true cross-covariance K_* (no population or large-n approximation) and is exact bookkeeping via Cauchy–Schwarz/Woodbury; it is also what we compute, exactly, in every experiment. An earlier draft additionally stated a population-level asymptotic approximating K_* via K^2/n (Bach, 2017, Lemma 2) to relate VC *in closed form to $\sum_i \lambda_i^2/(\lambda_i + \sigma^2)$; we removed it because we could not establish a rigorous error bound for the approximation (the distributional assumption on test points, the norm in which $\mathbb{E}[k_* k_*^\top] \approx K^2/n$ holds, and how that error behaves after the matrix inverse and trace are all unresolved), and, more concretely, it evaluates to $\lesssim \sigma^2$ for* every *spectrum under our normalization $K_q(x, x) = 1$—contradicting the measured* $\mathrm{VC} \approx 0.98$ *at Haar concentration (Figure 4). We therefore do not use it anywhere;* $\mathrm{VC} \to 1$ *at Haar concentration follows instead directly from Lemma 3's small-cross-covariance argument.*

Empirically, $\mathrm{VC}(K)$ ranges continuously from ≈ 0 at low spectral entropy (Lemma 2) to ≈ 1 at Haar concentration (Lemma 3); the expected calibration error $\mathrm{ECE}(K)$ (Appendix G) tracks a related but target-dependent spectral trend, discussed in Section 5.

4 The Universal Half: Spectrum Predicts Dequantization and Posterior Behavior

Ansatz families. To probe spectral universality we work with three families that span much of the variational-circuit literature (Figure 3, Appendix B): HE (hardware-efficient Ry/Rz rotations +

nearest-neighbor CNOT ring), MATCHGATE (single-qubit Rz + Ising XX/YY with distinct angles so that the circuit explores Spin($2n_q$) beyond the vacuum sector), and IQP (Hadamard sandwich enclosing diagonal Z and ZZ encodings) (Havlíček et al., 2019). Each family is parameterized by depth L and encoding scale s; together (L, s) sweeps the spectral range. Every kernel entry is estimated with the compute-uncompute primitive $U_\phi(x')U_\phi(x)^\dagger$.

Single-family concept experiment. Figure 4 (Appendix B) shows the four spectral panels for the HE family on $n_q = 6$ qubits across 24 (L, s) configurations on a synthetic target. As $S/\log n$ varies from 0.15 (deep constant collapse) to ≈ 1 (Haar limit), the dequantization score dq rises monotonically, the variance contraction ratio VC rises monotonically from 0.02 to 1, and the ECE drops from 0.38 to 0.07. Crucially, the predictive negative log likelihood (NLL) traces a clean U-shape with minimum ≈ 0.9 at $S/\log n \approx 0.91$, flanked by NLL above 10 at low entropy and 1.2 at high entropy. We dub this shape the *useful-hardness frontier*.

Cross-family universality. Repeating the sweep with all three families and 25 (L, s) points each gives the overlay in the top row of Figure 2: identical $S/\log n$ curves on the dq, ECE and variance-contraction panels, NLL minimum at $S/\log n \approx 0.85$–0.95 for every family, and a maximum vertical spread between any two families below 0.05 on the dq and contraction panels. The same picture holds at eight qubits, where the U-shape sharpens (Appendix B).

Why ECE alone is misleading. The ECE panel *decreases monotonically* with $S/\log n$, suggesting that highly entropic kernels are best calibrated. They are—but in the trivial sense of matching the prior: the NLL panel shows the same posterior to be uninformative and downstream useless. ECE is necessary but not sufficient; NLL is its informativeness-aware counterpart, and the two must be read together.

Comparison with alternative spectral diagnostics. Is $S/\log n$ a privileged axis or one of several reasonable choices? Appendix B compares it against seven established kernel-evaluation statistics on the same sweep. Several of them *correlate more strongly* with NLL (absolute Spearman: off-diagonal concentration 0.89, KTA 0.79 vs. $S/\log n$ 0.28), reflecting that NLL is *monotone* along those axes but *U-shaped* along ours. We adopt $S/\log n$ *not* because it maximizes correlation but because (i) it is label-free, so it serves as an a-priori diagnostic before fitting; (ii) it is normalized to $[0, 1]$, making cross-system comparison interpretable; and (iii) it exposes the useful-hardness frontier as a clean U-shape with an identifiable optimum—a feature absent from the monotone-correlated alternatives.

Scope of the universal-half claim. The three "universal" panels are not uniformly label-free: dq is a purely spectral quantity of K; VC is label-free at the population level (Proposition 1 depends only on K and the test-input distribution); ECE, computed from Gaussian credible intervals against realized test targets y_* (Appendix G), is genuinely label-dependent, so its consistent placement on the $S/\log n$ axis is an empirical finding, not a structural consequence of the parameterization. The non-trivial empirical content of the universal half is (i) that all three, plus the classical baselines below, are consistently organized by $S/\log n$ across very different gate sets once (L, s) is varied—which is what makes a single design axis viable—and (ii) that the predictive NLL still rides on the same axis, up to a target-dependent shift of its U-shape (Section 5). The genuinely predictive content of the framework is therefore the target-dependent NLL minimum and the downstream BO decision quality of Section 8.

Classical baselines collapse onto the same curves. A natural reviewer worry is that the spectral curves of Figure 2 are an artefact of using only quantum kernels: maybe any kernel family that spans the spectral range would look the same. To rule out this explanation we re-ran the entire sweep with five classical baselines (RBF, Matérn-3/2, Matérn-5/2, random Fourier features (Rahimi & Recht, 2007) at three widths, and a two-layer deep-kernel proxy at three widths). The result is striking: for each target separately, classical and quantum kernel *families* populate the *same* dq, ECE, and variance-contraction curves as a function of $S/\log n$—i.e. family identity does not matter—and, across targets, both kernel classes show the same NLL inversion (U-shape on synthetic, monotone increase on quantum-data) and the same ECE reversal (falling with entropy on synthetic, rising on quantum-data; Figure 1). The spectral diagnostic is therefore not a quantum-specific phenomenon:

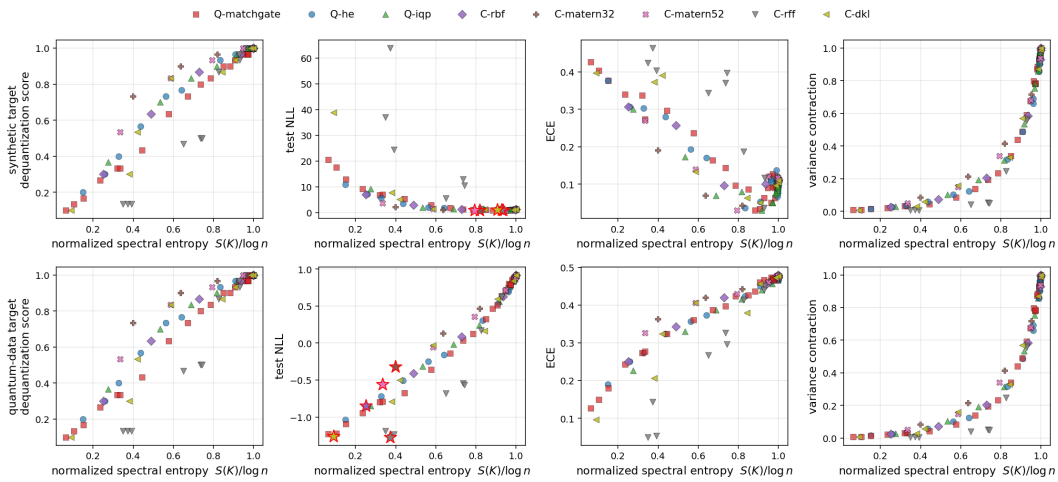


Figure 1: Classical kernels (squares/triangles/diamonds: RBF, Matérn-3/2, Matérn-5/2, RFF, deep-kernel) overlaid on quantum kernels (circles: HE, matchgate, IQP) at $n_q = 6$. All families fall on the same dq/ECE/VC curves as a function of normalized spectral entropy $S(K)/\log n$ *within each target* (column 1, 3, 4). The NLL panels (column 2) preserve the target-dependent shape: U-shape for the synthetic target (top), monotone increase for the quantum-data target (bottom). The spectral anatomy is kernel-agnostic.

it is a property of the GP posterior structure that the quantum and classical literatures both inherit. Practically, this means the recipe of Section 7 applies verbatim to classical-kernel GPs.

Real data and non-modular ansatze. Two further checks, reported in Appendix B, leave the picture unchanged. On two genuine regression benchmarks (*diabetes*, *california housing*) all four families (HE, matchgate, IQP, classical RBF) collapse onto the same dq/ECE/VC curves and place their NLL minimum at $S/\log n \approx 0.96$–0.99, with quantum kernels competitive with—but not significantly better than—RBF. And two *non-modular* ablation ansatze—random single-qubit Pauli rotations with *no* entangling gates, and Clifford+T layers alternating with data-dependent R_z encodings—fall on the same universal curves, so the collapse is not an artefact of the nearest-neighbour modular layer structure shared by the three main families.

5 The Target-Dependent Frontier

The kernel-only half of the story—dequantization and variance contraction—is target-independent, while calibration is empirically organized by the same spectral axis across our targets (Section 4). Predictive NLL, in contrast, explicitly couples kernel and target. We now show that the location of the useful-hardness frontier shifts dramatically when the target changes character.

Two targets. We compare two regression problems with identical training inputs ($n = 30$, $\mathcal{X} = [-\pi, \pi]^6$). The *synthetic target* is $y(x) = \sin(w^\top x) + 0.3\cos(2x_1) + 0.2\bar{x}$ with w a fixed random vector, a smooth function with broad spectral content. The *quantum-data target* is $y(x) = \langle 0^{n_q}|U_{\text{data}}(x)^\dagger(\frac{1}{n_q}\sum_q Z_q)U_{\text{data}}(x)|0^{n_q}\rangle$, where U_{data} is a separate $L = 3$ data-reuploading circuit unrelated to any of our kernel ansatze, so its band-limitation in the *data*-circuit eigenbasis does not by itself constrain the target's spectral mass $\beta = U^\top y$ in the unrelated *kernel* eigenbasis U of Definition 4—and, as we measure directly below, it indeed does not: the two bases are not aligned for this target.

Frontier shape inverts. For the synthetic target the NLL panel of Figure 2 reproduces a clear U-shape with minimum at $S/\log n \approx 0.9$; for the quantum-data target the NLL *increases* monotonically, its best value (≈ -1.2) reached at the lowest accessible entropy $S/\log n \approx 0.1$. The dq and

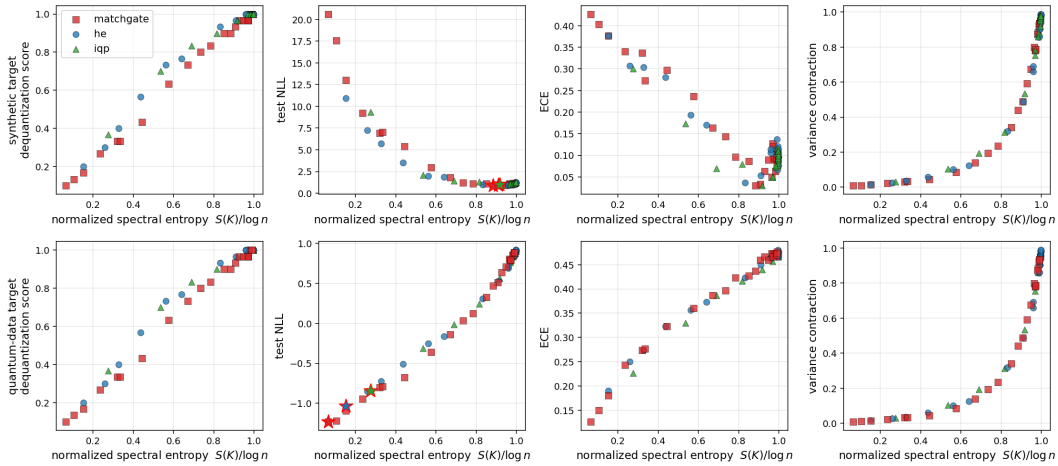


Figure 2: Spectral anatomy across two targets at $n_q = 6$. Top row: synthetic target with ± 1 s.d. cross-seed errorbars over 3 seeds (drawing both training inputs and ansatz parameters); the NLL minimum (stars) sits at $S/\log n \approx 0.9$, identifying a classical U-shape useful-hardness frontier. Bottom row: quantum-data target $y(x) = \langle\psi(x)|O|\psi(x)\rangle$ produced by a fixed $L = 3$ ansatz (single seed because the data-generating circuit is itself fixed); the NLL is monotonically *increasing* in $S/\log n$ and the optimal kernel lives at very low spectral entropy. The dq and VC panels, which are functions of K alone, are unchanged across targets; ECE, which uses the realized test targets, reverses sign along with NLL (falling with entropy on top, rising with entropy on bottom).

VC panels are unchanged by construction (both are functions of K alone), but ECE—which uses the realized targets—is not: it falls with entropy on the synthetic target and *rises* with entropy on the quantum-data target, mirroring the NLL reversal rather than staying fixed.

Interpretation: not spectral alignment, but the noise floor. We directly computed $\beta = U^\top y$ and $d_{\text{int}}(y, K)$ for the quantum-data target across the full 25-point HE grid ($S/\log n \in [0.15, 0.999]$), ruling out a spectral-alignment story: d_{int} stays at $[8.0, 17.3]$ of $n = 30$ throughout—largest, if anything, at the two lowest-entropy points (15.9, 13.2)—with no eigendirection ever exceeding 22% of the target's energy, even at the most rank-1-like config. A simpler mechanism follows from Theorem 2 itself: this target's variance (≈ 0.0034) is $15\times$ below the noise $\sigma^2 = 0.05$, so every measured β_i^2 (max ≈ 0.023) sits under the noise floor, and $\lambda_i^* = \max(0, \beta_i^2 - \sigma^2)$ drives nearly every eigenvalue to 0 regardless of *which* eigenvectors they are—not eigenbasis alignment, just the whole signal being sub-threshold (Appendix A). The synthetic target, by contrast, has $O(1)$ amplitude well above σ^2 and needs a high-rank kernel for its diffuse, above-noise content, paying the calibration cost of approaching Haar concentration. The useful-hardness frontier is a property of the *pair* (kernel, target), which the next result makes formal for the idealized (fixed-eigenbasis) setting.

Definition 4 (Target spectral content)**.** Let $K = U\Lambda U^\top$ be the eigendecomposition of K and let $\beta := U^\top y \in \mathbb{R}^n$ be the coordinates of the target vector in the kernel eigenbasis. The *target spectral mass* on the top-k eigendirections is $M_k(y; K) := \sum_{i=1}^k \beta_i^2 / \|\beta\|^2$.

Theorem 2 (NLL-optimal kernel spectrum)**.** *Fix $y \in \mathbb{R}^n$ and $\sigma^2 > 0$. Among PSD Gram matrices $K = \sum_i \lambda_i u_i u_i^\top$ with fixed eigenbasis $\{u_i\}_{i=1}^n$ and unconstrained eigenvalues $\lambda_i \geq 0$, the GP marginal log likelihood $\log p(y \mid K, \sigma^2)$ is maximized at*

$$\lambda_i^* \;=\; \max\!\big(0,\; \beta_i^2 - \sigma^2\big), \qquad \beta_i := u_i^\top y.$$

Under an additional trace budget $\sum_i \lambda_i \leq T$ with $T < \sum_i \max(0, \beta_i^2 - \sigma^2)$, KKT stationarity replaces this water-filling solution by the positive root of a per-index quadratic $\eta t^2 + t - \beta_i^2 = 0$ in $t = \lambda_i^* + \sigma^2$, whose $\eta \to 0^+$ expansion recovers the unconstrained optimum up to a β_i-dependent correction rather than a uniform rescaling (Appendix H).

Corollary 1 (Effective rank matches target intrinsic dimension)**.** *Let* $\hat{\beta}_i := \beta_i^2/\|\beta\|^2$ *and define the* intrinsic dimension *of* y *in the kernel eigenbasis as* $d_{\text{int}}(y, K) := 1/\sum_i \hat{\beta}_i^2 \in [1, n]$*. In the high-SNR regime (*$\beta_i^2 \gg \sigma^2$ *for* $i \in \text{supp}\{\lambda_i^*\}$*), the NLL-optimal kernel* K^* *of Theorem 2 satisfies*

$$r_{\text{eff}}(K^*) \;=\; \frac{(\sum_i \lambda_i^*)^2}{\sum_i (\lambda_i^*)^2} \;\approx\; \frac{1}{\sum_i \hat{\beta}_i^2} \;=\; d_{\text{int}}(y, K^*),$$

where the first equality is exact (r_{eff} *evaluated at* $\lambda_i^* = \max(0, \beta_i^2 - \sigma^2)$*); the second is the high-SNR approximation itself, exact only as* $\sigma^2/\beta_i^2 \to 0$ *on the active set, since* $\lambda_i^* = \beta_i^2 - \sigma^2 \neq \beta_i^2$ *for finite* σ^2*. and the optimal spectral entropy satisfies* $s^*(K) := S(K^*)/\log n \in \left[\log d_{\text{int}}/\log n,\ 1\right]$ *with the lower bound attained at degenerate (top-*d_{int}*) spectra. Band-limited targets (*$d_{\text{int}} \sim 1$*) thus prefer kernels with* $s \to 0$*, while broadband targets (*$d_{\text{int}} \sim n$*) require* $s \to 1$*.*

The corollary unifies the U-shape on synthetic data and the monotone behaviour on quantum-data targets within a single qualitative story about d_{int}. It should be read as intuition rather than a tight quantitative prediction for our sweeps: Theorem 2 optimizes *unconstrained* eigenvalues over a *fixed* eigenbasis U, whereas the M2 sweep instead varies circuit parameters (L, s), which change the eigenbasis itself and are restricted to whatever spectra the ansatz family can realize—a different, non-convex optimization problem over a restricted set of K's, not the one Theorem 2 solves. Consequently the exact identity $r_{\text{eff}}(K^*) = d_{\text{int}}(y, K^*)$ of Corollary 1 is not expected to, and does not, hold quantitatively on the swept optimum (Table 2 below). We do *not* claim the corollary's looser interval bound on s^* transfers either—a bound derived from the same idealized model has no automatic claim to validity once the model demonstrably does not describe the real optimization; what we report below is only the qualitative direction (larger targets need higher entropy) and a descriptive, not confirmatory, comparison against the interval (Table 2). Estimating d_{int} a priori is non-trivial; in practice we read it off the spectrum of a single trial kernel as part of the recipe in Section 7.

Scaling. At $n_q = 8$ the U-shape sharpens and its minimum migrates to $S/\log n \approx 0.98$: larger Hilbert spaces require more spectral entropy to populate informative directions, so the useful-hardness window contracts as n_q grows. The measured d_{int} (computed directly from $\beta = U^\top y$ at each swept optimum, not derived from Corollary 1) also grows with n_q, from 5.2 at $n_q = 6$ to 12.2 at $n_q = 8$. We note this alongside the observed optima 0.91 and 0.99 as a descriptive fact, not as confirmation of Corollary 1's mechanism: since the swept optimization does not match the corollary's idealized setting (Appendix B below), the fact that these observed values happen to fall inside $[\log d_{\text{int}}/\log n, 1]$ is not evidence the corollary's water-filling argument is what produces them.

6 Hardware Validation on IBM Heron

A diagnostic that exists only in simulation is unconvincing. We validate the spectral anatomy on a real superconducting device.

Setup. We use `ibm_aachen`, a 156-qubit IBM Heron-class processor. For each of 24 frontier configurations (8 HE, 8 matchgate, 8 IQP), we draw $n = 8$ inputs uniformly from $[-\pi, \pi]^4$, build the compute-uncompute overlap circuit $U_\phi(x)U_\phi(x')^\dagger$, and sample 4096 shots per pair, comparing exact statevector simulation, Aer with shot noise only, and `ibm_aachen` with *no error mitigation*. As a rough normalization we also report $z := |\Delta S|/\sigma_{\text{sim}}$, where σ_{sim} is the noiseless statevector's standard deviation of $S/\log n$ over five random seeds of the training inputs and ansatz parameters at the same (ansatz, L, s). This is *not* a measurement of hardware or shot noise—it is configuration-to-configuration spread under resampled randomness, a different source of variation—so z should be read as a coarse same-order-of-magnitude check, not a calibrated significance test; the 5-seed same-configuration hardware reruns of Appendix J are the more direct evidence for run-to-run hardware stability.

Result. The diagnostic transfers with median absolute error 0.032 in $S/\log n$ and mean 0.052 (Table 1), and the story is *family-stratified*: matchgate and IQP transfer with $\leq 9\%$ worst-case error and $\bar{z} \leq 1.1$, whereas HE degrades to 30% at depth $L \geq 3$, where the transpiled compute-uncompute circuit is longest. This is the expected signature of gate-noise accumulation in deep HE

Table 1: Hardware transfer summary on `ibm_aachen` (156-qubit Heron), $n_q = 4$, $n = 8$, 4096 shots, no error mitigation. Per-family aggregate over 8 frontier configurations each. Matchgate and IQP match the noiseless simulator within 5% mean absolute error; HE is the most noise-sensitive family because its data-encoding layers compose with CNOT entanglers, lengthening the transpiled circuit. 15 of 24 points have $z = |\Delta S|/\sigma_{\text{sim}} < 2$.

family	n_{pts}	$\lvert\Delta S\rvert_{\max}$	$\lvert\Delta S\rvert_{\text{mean}}$	$\lvert\Delta S\rvert_{\text{med.}}$	$\bar{z}$	#($z < 2$)
HE	8	0.300	0.101	0.070	3.2	2/8
matchgate	8	0.053	0.024	0.024	1.1	6/8
IQP	8	0.093	0.030	0.020	0.6	7/8
overall	24	0.300	0.052	0.032	1.6	15/24

encodings—a hardware limitation, not a failure of the diagnostic: the eight HE points' hardware-vs.-exact ranking is strongly but not exactly preserved (Spearman $\rho = 0.93$, Kendall $\tau = 0.79$; Table 9), with the depth-$L \geq 3$ outlier and one adjacent pair swapping rank.

Noise interpretation. The bias is not random: low-entropy configurations shift *upward* (mean $+0.054$) and near-Haar ones shift slightly *downward* (mean -0.074). Appendix F traces this to depolarizing noise adding a rank-1 component to the Gram matrix (Lemma 4), which *lowers* s near Haar concentration and barely moves it near constant collapse, while shot/readout noise flattens the spectrum at low s—together predicting the observed pattern (scatter plot: Appendices B, K).

7 A Practical Recipe

Algorithm 1 (Appendix E) turns the spectral anatomy into a diagnostic: compute one Gram matrix, read off $S/\log n$, r_{eff}/n, and return a regime. Below 0.1: **constant collapse**, use only if band-limited. High entropy alone is not grounds for rejection—Section 8's best BO surrogate and Figure 6's real-data winners both sit at $S/\log n \approx 0.95$–0.99—so above 0.95 we only flag **near-Haar** when a held-out split *also* shows the posterior variance failing to contract (VC ≥ 0.9), and recommend reducing L or s only then; otherwise the configuration is **useful**, and a held-out split decides between K_q and an r_{eff}-feature Nyström/RFF surrogate. Cost: one $O(n^2)$ Gram evaluation plus an $O(n^3)$ eigendecomposition, applying to NISQ devices as well as simulators (Section 6).

8 Downstream Validation: Bayesian Optimization

We now test whether the spectral regime predicts *decision quality* in downstream Bayesian optimization (BO), comparing four GP surrogates on the two objectives of Section 5—HE kernels at low/mid/high spectral entropy (Q-CONSTANT ≈ 0.18, Q-SWEET ≈ 0.99, Q-HAAR ≈ 1.00) and a classical CLS-RBF baseline, each run for $T = 25$ UCB-BO (Srinivas et al., 2010) iterations—and tracking the best objective found so far (Figure 11, Appendix B; 8 seeds quantum-data, 6 synthetic, ± 1 s.e.). On *synthetic*, Q-HAAR is clearly worst (-1.31, sem 0.06) while the other three are indistinguishable (-1.44, -1.41, -1.40; sem ≈ 0.09), so a non-contracting posterior does starve UCB on this broadband target, as predicted. On *quantum-data*, however, Q-SWEET leads (-0.104, sem 0.012) but Q-HAAR is *not* the worst of the rest (Q-HAAR -0.092, Q-CONSTANT -0.086, CLS-RBF -0.085; all indistinguishable, sem ≈ 0.01)—the spectral prediction holds on the broadband target but not on the band-limited one.

9 Conclusion

A single spectral quantity—the normalized entropy of the kernel Gram matrix—tracks dequantization difficulty, posterior calibration and variance contraction across three quantum ansatz families and several classical baselines, and transfers to IBM Heron hardware with median absolute error 0.032 and no error mitigation. The useful-hardness frontier itself is target-dependent: smooth targets prefer high-entropy kernels, band-limited quantum-data targets low-entropy ones (scope and limitations: Appendix A).

Reproducibility statement

All theoretical claims are stated with their assumptions in Section 3 and proved in full in Appendix H. Circuit families, kernel constructions, data sets, entropy estimators, and all hyperparameters used in the simulator experiments are specified in Appendices G and I; the hardware runs, including backend, qubit counts, shot budgets, and the absence of error mitigation, are detailed in Appendix K.

References

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of machine learning research*, 18(21):1–38, 2017.

Meng-Han Chen, Chao-Hua Yu, Jian-Liang Gao, Kai Yu, Song Lin, Gong-De Guo, and Jing Li. Quantum algorithm for gaussian process regression. *Physical Review A*, 106(1):012406, 2022.

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

Jun Dai and Roman V Krems. Quantum gaussian process model of potential energy surface for a polyatomic molecule. *The Journal of Chemical Physics*, 156(18):184802, 2022. doi:10.1063/5.0088821.

Zhongxiang Dai, Gregory Kang Ruey Lau, Arun Verma, Yao Shu, Bryan Kian Hsiang Low, and Patrick Jaillet. Quantum bayesian optimization. *Advances in neural information processing systems*, 36:20179–20207, 2023.

Ahmad Farooq, Cristian A Galvis-Florez, and Simo Särkkä. Quantum-assisted hilbert-space gaussian process regression. *Physical Review A*, 109(5):052410, 2024.

Cristian A Galvis-Florez, Ahmad Farooq, and Simo Särkkä. Quantum-assisted gaussian process regression using random fourier features. In *2025 IEEE International Conference on Quantum Software (QSW)*, pp. 22–27. IEEE, 2025.

Diego Garcıa-Martın, Martin Larocca, and Marco Cerezo. Deep quantum neural networks form gaussian processes. *arXiv preprint arXiv:2305.09957*, 2023.

Filippo Girardi and Giacomo De Palma. Trained quantum neural networks are gaussian processes. *Communications in Mathematical Physics*, 406(4):92, 2025.

Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009.

Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.

Junpeng Hu, Jinglai Li, Lei Zhang, and Shi Jin. A quantum gradient descent algorithm for optimizing gaussian process models. *Mathematical Models and Methods in Applied Sciences*, 36(08): 1715–1737, 2026. doi:10.1142/S0218202526420042.

Jonas Jäger, Paolo Braccia, Pablo Bermejo, Manuel G Algaba, Diego García-Martín, and M Cerezo. Provable and scalable quantum gaussian processes for quantum learning. *arXiv preprint arXiv:2605.00099*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. *Advances in Neural Information Processing Systems*, 34:12661–12673, 2021.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Dominic Lowe, MS Kim, and Roberto Bondesan. Assessing quantum advantage for gaussian process regression. *arXiv preprint arXiv:2505.22502*, 2025.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018.

Anderson Melchor Hernandez, Filippo Girardi, Davide Pastorello, and Giacomo De Palma. Quantitative convergence of trained quantum neural networks to a gaussian process. *Annales Henri Poincaré*, pp. 1–57, 2025. doi:10.1007/s00023-025-01631-6.

Matthew Otten, Imene R Goumiri, Benjamin W Priest, George F Chapline, and Michael D Schneider. Quantum machine learning using gaussian processes with performant quantum kernels. *arXiv preprint arXiv:2004.11280*, 2020.

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

Frederic Rapp and Marco Roth. Quantum gaussian process regression for bayesian optimization. *Quantum Machine Intelligence*, 6(1):5, 2024.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

Ewin Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st annual ACM SIGACT symposium on theory of computing*, pp. 217–228, 2019.

Supanut Thanasilp, Samson Wang, Marco Cerezo, and Zoë Holmes. Exponential concentration in quantum kernel methods. *Nature communications*, 15(1):5200, 2024.

Joel J Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5):052325, 2016.

Zhikuan Zhao, Jack K Fitzsimons, and Joseph F Fitzsimons. Quantum-assisted gaussian process regression. *Physical Review A*, 99(5):052331, 2019a.

Zhikuan Zhao, Jack K Fitzsimons, Michael A Osborne, Stephen J Roberts, and Joseph F Fitzsimons. Quantum algorithms for training gaussian processes. *Physical Review A*, 100(1):012304, 2019b.

A Discussion and Limitations

What the spectral anatomy does not say. We do not claim that the useful-hardness frontier corresponds to a regime of genuine *computational* quantum advantage over the best classical algorithm. Configurations with $S/\log n \approx 0.9$ are characterized by full-rank spectra, but the entries of those spectra may still admit polynomial-time classical approximation (Lowe et al., 2025). Our claim is more modest: spectrum predicts *behavior*—calibration, contraction, classical-RFF difficulty—without itself certifying complexity-theoretic hardness.

Band-limitation does not transfer eigenbases, and does not need to. Section 5 originally conjectured that the quantum-data target's preference for low-entropy kernels reflects alignment between the data-generating circuit's eigenbasis and the kernel's eigenbasis (small d_{int}). Direct measurement of $\beta = U^\top y$ and $d_{\text{int}}(y, K)$ across the full 25-point HE grid rules this out: d_{int} stays at 8.0–17.3 (out of $n = 30$) across the entire entropy range—if anything, the two *lowest*-entropy configurations have the *largest* d_{int} (15.9 and 13.2)—and no eigendirection ever captures more than 22% of the target's energy: the two eigenbases are not aligned for this target. Corollary 1 would, on its own idealized terms, predict this misalignment should force the frontier back toward high entropy; that it does not is itself evidence that the corollary's mechanism is not what is operating here. Instead, Section 5 identifies a simpler explanation directly from Theorem 2: this particular quantum-data target's variance (≈ 0.0034) sits well below the fixed observation noise $\sigma^2 = 0.05$, so the theorem's own optimal solution drives essentially every eigenvalue toward 0 irrespective of eigenbasis alignment. Whether this noise-floor mechanism also explains the earlier-observed frontier shift with n_q (Table 2, a different, broadband synthetic target where the signal is well above the noise floor and d_{int} genuinely governs the optimum) is a separate question we have not resolved; the two targets may simply sit in different regimes of the same underlying water-filling picture.

Beyond the three ansatz families. The matchgate family is the case where rigorous QGP scalability has recently been established (Jäger et al., 2026); our results suggest that even within that family the useful-hardness frontier is target-dependent. Extending the spectral analysis to other symmetry classes—in particular Clifford+T, IQP with non-NN couplings, and Hamiltonian variational ansätze beyond free-fermion—is an immediate direction.

Pareto frontier of regularization. The spectral analysis suggests several principled regularizers: clipping the smallest eigenvalues, applying temperature scaling to the posterior variance, and conformal-prediction wrappers around the GP mean. We sketch them in Appendix M but defer a full empirical study to follow-up work.

From diagnosis to design. The natural next step is to use the spectral diagnostic as a training signal: optimize ansatz parameters or depth so that $S/\log n$ matches the target-implied optimum. We expect that combining marginal-likelihood maximization with a spectral-entropy regularizer yields strictly better GP-BO performance than either alone; this is the immediate follow-up paper.

Statistical disclosure. We are explicit about what our experiments do and do not establish. The main M1/M2 sweeps use a single seed per (ansatz, L, s) config and report per-config scatter. Figure 4 reports cross-seed error bars (± 1 s.d. over 3 seeds drawing both training inputs and ansatz parameters) on each (L, s) aggregate, together with Spearman rank correlations of each spectral panel with $S/\log n$. The BO study (Section 8) uses 8 seeds per (surrogate, target) cell on the quantum-data target and 6 seeds on the synthetic target. The between-surrogate ordering on the synthetic target (Q-HAAR worst, the other three indistinguishable) is consistent across seeds; on the quantum-data target only Q-SWEET separates from the pack, and the gap to the runner-up is on the order of one standard error of the mean, so we cannot claim bulletproof statistical significance there, nor that Q-HAAR is uniformly the worst surrogate. The real-data evaluation (Figure 6) is a sanity check on *two* small benchmarks (diabetes, california housing) and shows that quantum kernels are *competitive with*, not significantly better than, an RBF baseline; the contribution we claim is the spectral diagnostic, not a quantum advantage on these tasks. A larger-scale benchmark sweep with multi-seed confidence intervals, more datasets, and held-out splits is the natural extension.

B Additional Empirical Results

B.1 Circuit diagrams

B.2 Single-family concept sweep

B.3 Alternative spectral diagnostics

A natural question is whether $S/\log n$ is a privileged axis or just one of several reasonable choices. We compute seven established kernel-evaluation statistics on the same M1 sweep (Figure 5): the

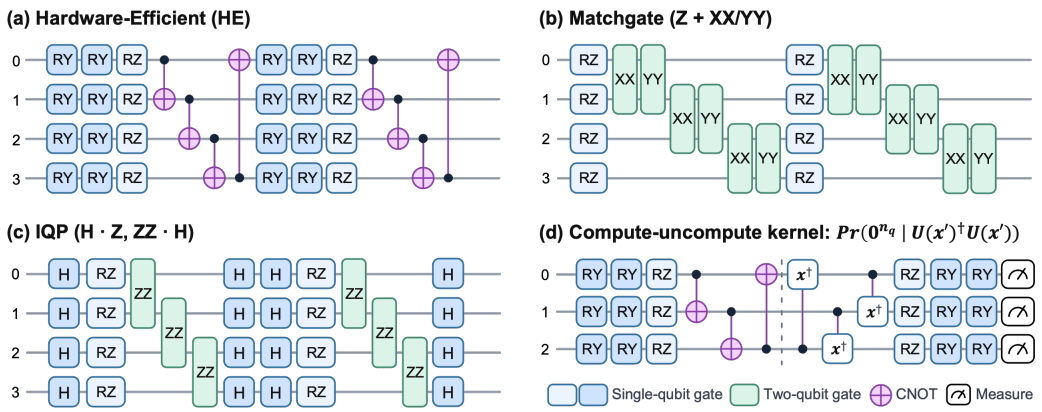


Figure 3: Quantum circuits used in this work, shown at depth $L = 2$ and $n_q = 4$ for the three ansatz families and a compact $L = 1$, $n_q = 3$ instance for the kernel primitive. (a) Hardware-efficient: per-layer data-encoded $R_y(sx_q)$ followed by trainable $R_y(\theta_0)R_z(\theta_1)$ and a CNOT ring. (b) Matchgate: trainable $R_z(\theta_0)$ then per-pair $\mathrm{IsingXX}(s(x_q + x_{q+1})/2)$ and $\mathrm{IsingYY}(\theta_1)$; the two distinct angles break particle-number conservation, which is essential for the kernel to depend on the input. (c) IQP: Hadamard sandwich enclosing $R_z(sx_q c_q^{(\ell)})$ and $\mathrm{IsingZZ}(sx_q x_{q+1} d_{q,q+1}^{(\ell)})$. (d) Compute-uncompute primitive: every kernel entry $K_q(x, x') = |\langle 0|U_\phi(x')^\dagger U_\phi(x)|0\rangle|^2$ is estimated as the all-zero probability after running $U_\phi(x)$ followed by $U_\phi(x')^\dagger$.

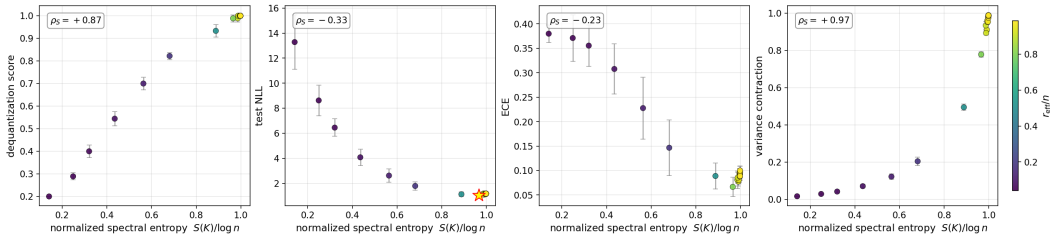


Figure 4: Single-family concept verification on the hardware-efficient ansatz ($n_q = 6$, $n = 30$, synthetic target, 3 seeds per configuration). Across 24 ($L \in \{1, 2, 3, 5, 8, 12\}$, $s \in \{0.1, 0.5, 1, 2\}$) configurations a single spectral quantity $S(K)/\log n$ predicts: (left) dequantization difficulty, (middle-left) test NLL with a U-shape whose minimum (star) identifies the useful-hardness frontier, (middle-right) regression ECE, (right) variance contraction. Vertical bars are ± 1 s.d. over the three seeds; ρ_S is the Spearman rank correlation between the metric and $S/\log n$.

effective rank ratio r_{eff}/n, Bach's degrees of freedom d_σ/n (Bach, 2017), log condition number $\log_{10} \kappa(K)$, kernel–target alignment (Cristianini et al., 2001), centered KTA (Kornblith et al., 2019), off-diagonal kernel concentration $c(K)$, and the trace-truncation rank $\mathrm{dq}_{0.01}$. Several alternatives *correlate more strongly* with NLL than $S/\log n$ does (absolute Spearman: off-diagonal concentration 0.89, KTA 0.79, $\mathrm{dq}_{0.01}$ 0.64 vs. $S/\log n$ 0.28), reflecting that NLL is *monotone* along those axes but *U-shaped* along $S/\log n$. On the same data, Spearman rank correlations with test NLL are -0.28 ($S/\log n$), -0.28 (r_{eff}/n), -0.64 (dq), $+0.66$ (ECE) and -0.18 (VC); none dominates as a monotone predictor, but only $S/\log n$ exposes the U-shape with a target-dependent minimum while remaining label-free and normalized to $[0, 1]$. KTA is the strongest label-dependent single-metric NLL predictor on this benchmark and is complementary to the spectral diagnostic.

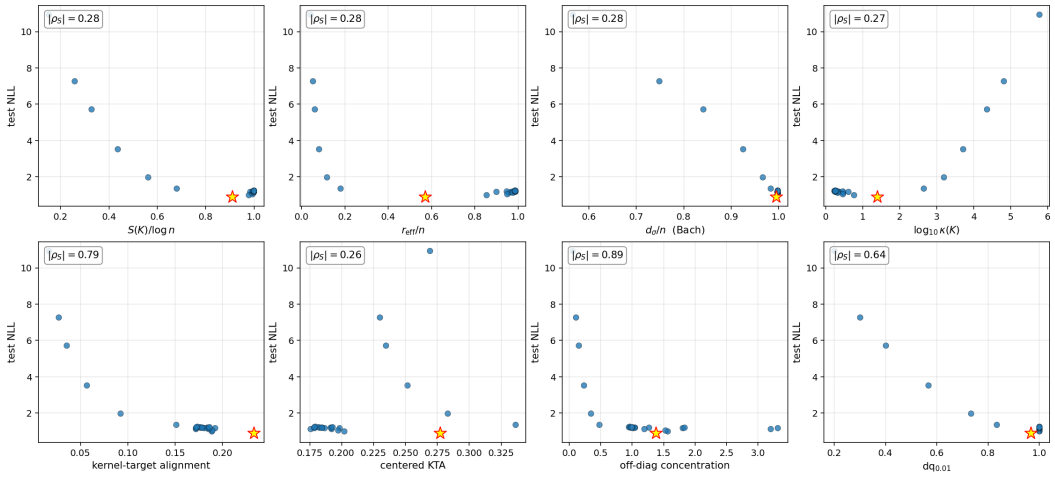


Figure 5: Eight candidate spectral diagnostics, plotted against test NLL on the M1 sweep ($n_q = 6$, HE, 24 configurations). $|\rho_S|$ is the absolute Spearman correlation with NLL. Off-diagonal concentration and KTA are most strongly correlated with NLL because NLL is approximately monotone along those axes; the entropy-family axes $S/\log n$, r_{eff}/n, d_σ/n have lower Spearman but display the U-shape that exposes the useful-hardness sweet spot (star).

B.4 Real-data benchmarks

Real-data benchmarks confirm the picture. The synthetic and quantum-data targets of Section 5 are designed to span opposite ends of the target-intrinsic-dimension spectrum. To check that the universal-half and target-dependent-half predictions transfer to genuine ML benchmarks, we reran the M2 sweep on two real regression datasets from `scikit-learn`: *diabetes* (442 medical samples, 10 features) and *california housing* (200 random samples of 20 640, 8 features). Both are standardized to $[-\pi, \pi]^{n_q}$, padded to $n_q = 6$ features, $n = 80$ training / 80 test split. Figure 6 shows that all four families (HE, matchgate, IQP, classical RBF) collapse onto the same dq/ECE/VC curves and place their NLL minimum at $S/\log n \approx 0.96$–0.99, consistent with the broadband synthetic-target prediction. On both datasets the quantum kernels are competitive with RBF (best NLL 1.26 for IQP vs. 1.30 for RBF on california; 1.31 for HE vs. 1.29 for RBF on diabetes), confirming that the spectral diagnostic correctly identifies the regime in which quantum kernels are useful but does not claim a quantum-kernel *advantage* on these particular tasks.

B.5 Ablation: non-modular ansatze

Ablation: collapse persists under non-modular ansatze. The three quantum families used so far (HE, MATCHGATE, IQP) all share a NN, modular layer structure. To check that the universal collapse is not a side-product of this commonality, we run two non-modular ablation ansatze: RANDPAULI-NOENT, which applies only random single-qubit Pauli rotations with *no entangling gates* at all, and CLIFFORDT, which alternates random Clifford layers with data-dependent R_z encodings. Figure 8 shows that both ablations fall onto the same dq, ECE, and variance-contraction curves as the modular families. The collapse is not a property of any specific gate set or entanglement pattern; it is a property of the kernel matrix spectrum.

B.6 Scaling to $n_q = 8$

Scaling. Figure 9 repeats the synthetic-target analysis at $n_q = 8$ qubits with 30 points per family. The U-shape becomes sharper and shifts right: the best NLL now sits at $S/\log n \approx 0.98$. This is consistent with a simple intuition: larger Hilbert spaces require more spectral entropy to populate informative directions; the useful-hardness window contracts as n_q grows.

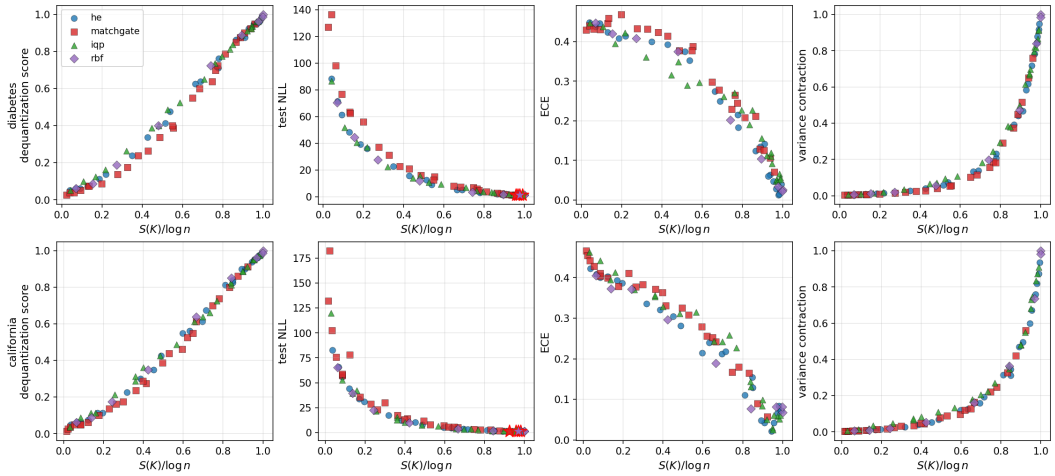


Figure 6: Spectral anatomy on two real regression benchmarks (*diabetes* top, *california housing* bottom). Three quantum ansatz families (he/matchgate/iqp circles) and an RBF baseline (purple diamonds) all collapse onto the same dq/ECE/VC curves; the NLL sweet spot (stars) lives at $S/\log n \approx 0.96$–0.99 for every family, matching the broadband target regime of Section 5. Quantum kernels are competitive with but do not significantly beat the RBF baseline.

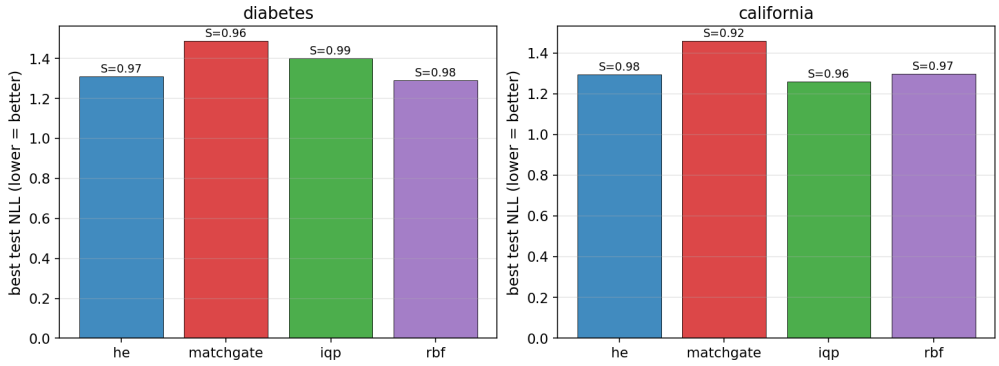


Figure 7: Best test NLL per kernel family on the two real-data benchmarks. Annotations show the $S/\log n$ value at the best configuration: every winner lives in the $S/\log n \in [0.92, 0.99]$ band, consistent with the broadband-target regime of Section 5. Quantum kernels are competitive with the RBF baseline (best HE is 1.30 vs. RBF 1.29 on diabetes; best IQP is 1.26 vs. RBF 1.30 on california) but do not significantly outperform it.

What Table 2 does and does not show. Theorem 2/Corollary 1 optimize eigenvalues freely over a *fixed* eigenbasis; the M2 sweep instead varies circuit parameters, which change the eigenbasis itself under ansatz-family constraints. That these are different optimization problems already shows up quantitatively: measured $r_{\text{eff}}(K^*)$ is 16.7 and 39.7 against measured $d_{\text{int}}(y, K^*)$ of 5.2 and 12.2 (both computed directly from $\beta = U^\top y$ at the swept optimum, not derived from the corollary), roughly 3–3.25$\times$ larger than the corollary's exact identity $r_{\text{eff}}(K^*) = d_{\text{int}}(y, K^*)$ would require. Given that the identity already fails, we do *not* treat the corollary's looser interval bound $s^*(K) \in [\log d_{\text{int}}/\log n,\ 1]$ as validated either—there is no a priori reason a bound derived from the same idealized model should hold once the model demonstrably does not describe the real optimization. Table 2 instead just reports the raw comparison as a descriptive fact: the observed s^* (0.91 at $n_q = 6$, 0.99 at $n_q = 8$) happens to fall inside that interval ($[0.49, 1.00]$ and $[0.68, 1.00]$) in both rows we have, which is consistent with, but not evidence for, the corollary's water-filling mechanism actually operating in the swept ansatz family.

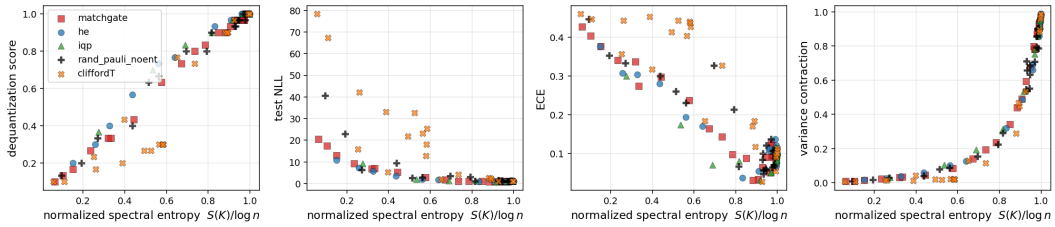


Figure 8: Ablation. RAND-PAULI-NOENT (black $+$, NO entangling gates) and CLIFFORDT (orange $\times$) overlaid on HE/MATCHGATE/IQP. The universal spectral curves are unchanged. The NLL panel shows mild scatter for CLIFFORDT at low S (sensitivity to random Clifford choice), but the average trend matches.

Table 2: Descriptive comparison at the NLL-optimal configuration of each (n_q, ansatz) pair on the synthetic target: measured $r_{\text{eff}}(K^*)$ and $d_{\text{int}}(y, K^*)$ (both computed directly from the swept optimum, not from Corollary 1), and where the observed s^* falls relative to the corollary's interval $[\log d_{\text{int}}/\log n,\ 1]$. Neither the corollary's exact identity nor its interval bound should be read as validated by this table—see text.

ansatz	n_q	best (L, s)	$r_{\text{eff}}(K^*)$	$d_{\text{int}}(y, K^*)$	observed s^* / interval
he	6	$(3, 0.30)$	16.7	5.2	0.91 / $[0.49, 1.00]$
he	8	$(1, 1.00)$	39.7	12.2	0.99 / $[0.68, 1.00]$

B.7 SIMULATOR-VERSUS-HARDWARE SCATTER

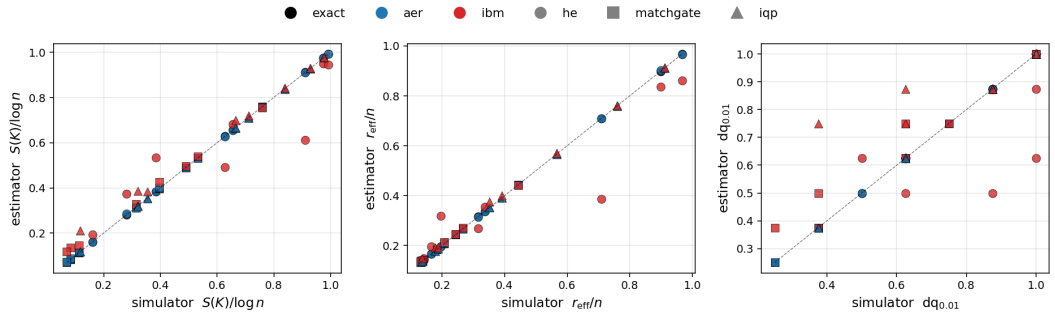


Figure 10: Hardware validation of the spectral diagnostic on `ibm_aachen` (156-qubit Heron, $n_q = 4$, $n = 8$, 4096 shots, no error mitigation). 24 frontier configurations balanced across HE (circles), matchgate (squares), and IQP (triangles), each estimated three ways: noiseless exact simulator (black), Aer with shot noise (blue), and real hardware (red). Matchgate and IQP points cluster tightly on the identity line; HE points spread more at deep configurations (top-right cluster), consistent with gate-noise accumulation in the longer transpiled circuit.

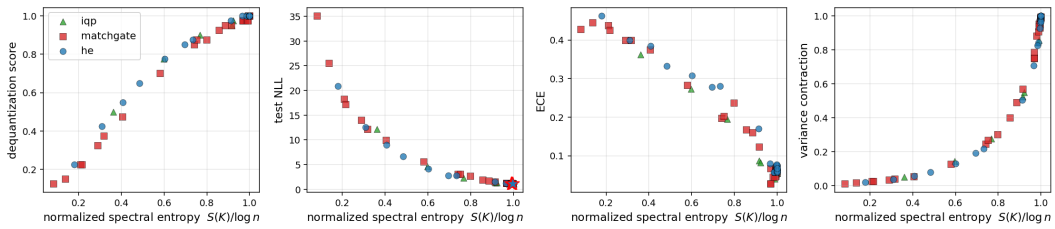


Figure 9: Scaling at $n_q = 8$ qubits, synthetic target, three ansatz families, 90 total configurations. The qualitative picture is preserved but the useful-hardness frontier sharpens: the NLL minimum migrates to $S/\log n \approx 0.98$ and the U becomes deeper. Larger Hilbert spaces shrink the useful spectral window.

B.8 Downstream Bayesian optimization

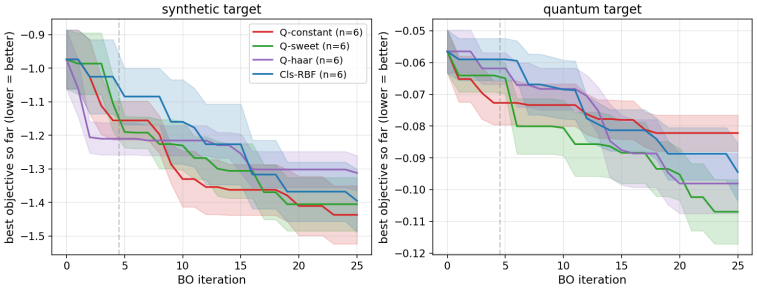


Figure 11: BO best-objective-so-far curves, shaded ± 1 s.e. (Left) Smooth synthetic objective (6 seeds): the three non-Haar surrogates are indistinguishable and Q-HAAR lags. (Right) Quantum-data objective (8 seeds): Q-SWEET pulls ahead from iteration $\approx$15; the other three surrogates, Q-HAAR included, are statistically indistinguishable from each other on this objective.

C Extended Related Work

Quantum GPs and HHL-based speedups. The seminal QGP algorithms (Zhao et al., 2019a;b) construct posterior mean and variance estimates via quantum linear systems (Harrow et al., 2009). Chen et al. (2022) relax the oracular kernel assumption via coherent-state amplitude estimation; Hu et al. (2026) provide a quantum gradient method for hyperparameter training; Farooq et al. (2024) and Galvis-Florez et al. (2025) exploit reduced-rank approximations to escape the dense-Hilbert-space regime.

Dequantization and lower bounds. Tang (2019) initiated the dequantization program. Lowe et al. (2025) prove that the condition number of the kernel matrix in many QGP settings grows at least linearly with n, implying at most polynomial quantum advantage. This neutralizes the exponential speedups conjectured by the HHL line and relocates the question from "is there a speedup" to "what spectral structure makes a kernel non-trivially quantum."

Quantum-NNGP theory. A parallel theoretical line establishes that quantum neural networks in suitable limits converge to Gaussian processes (Garcıa-Martın et al., 2023; Girardi & De Palma, 2025; Melchor Hernandez et al., 2025), providing the quantum analogue of the classical NNGP correspondence (Lee et al., 2018; Matthews et al., 2018). Jäger et al. (2026) construct a family of scalable, provable QGPs based on matchgate symmetry, demonstrating phase-diagram learning on up to 100 qubits.

Hardware-side QGPs and applications. Otten et al. (2020) run a five-qubit QGP regression on the IBM Boeblingen device, and Rapp & Roth (2024) demonstrate quantum kernel Bayesian optimization on `ibmq_montreal`. On the application side, Dai & Krems (2022) apply a quantum GP model to learn the potential-energy surface of a polyatomic molecule, illustrating the chemistry use case our spectral diagnostic targets.

Kernel concentration. Quantum kernels constructed from deep random circuits suffer exponential concentration of measure (Thanasilp et al., 2024; Kübler et al., 2021), the kernel analogue of the barren plateau phenomenon in variational quantum algorithms (McClean et al., 2018). Concentration is typically diagnosed by the variance of off-diagonal kernel entries; we will see in Section 3 that this variance is a coarse proxy for the spectral statistics that more directly control downstream behavior.

Gap. Existing QGP literature treats kernel-dequantizability and kernel-concentration as separate phenomena tied to different sub-fields (quantum complexity vs. trainability). No prior work analyzes them via a single spectral framework, and no prior work isolates the role of the target in selecting the useful spectral regime. The present paper supplies both.

D THE TWO REFERENCE PATHOLOGIES

Lemma 2 (Constant collapse)**.** *If $K = c\mathbf{1}\mathbf{1}^\top + \delta E$ is PSD with $\|E\|_{\mathrm{op}} \le 1$, $c > 0$, and $0 < \delta \le c/2$, then*

$$r_{\mathrm{eff}}(K) = 1 + O(\delta/c), \quad S(K) = O\big((\delta/c)\log(c/\delta)\big).$$

(Shannon entropy is not Lipschitz at a rank-1 spectrum: the bound is $\Theta(\delta/c)$ up to the extra $\log(c/\delta)$ factor, not $O(\delta/c)$ itself.) Suppose in addition the test–train cross-covariance satisfies $k_* = c\mathbf{1} + \delta' e$ for some $\|e\|_\infty \le 1$, $0 \le \delta' \le c/2$ (i.e. the kernel is constant-like at the test point too). Then for any fixed *$\sigma^2 > 0$, the σ^2-regularized GP posterior satisfies*

$$\mu(x_*) = \frac{nc}{nc+\sigma^2}\,\bar{y} + O_\sigma\big(\max(\delta,\delta')/c\big), \qquad v(x_*) = K_q(x_*,x_*) - \frac{nc^2}{nc+\sigma^2} + O_\sigma\big(\max(\delta,\delta')/c\big)$$

as $\delta, \delta' \to 0$ (the implicit constants depend on σ^2). Note the leading term is $\frac{nc}{nc+\sigma^2}\bar{y}$, a shrunk *empirical mean, not $\bar{y}$ itself—the shrinkage factor $nc/(nc+\sigma^2) < 1$ is a fixed regularization bias that does* not *vanish as $\delta, \delta' \to 0$ for fixed σ^2 (it only $\to 1$ as $\sigma^2 \to 0$ or $n \to \infty$). The variance is x_*-independent up to the stated $O_\sigma(\cdot)$ term. Both conclusions require σ^2 fixed away from 0: for $\sigma^2 = 0$ exactly, $\mu(x_*)$ need not converge to* any *multiple of $\bar{y}$ as $\delta \to 0$, because K^{-1} amplifies the residual signal along the kernel's near-null direction by a factor $\sim 1/\delta$, which can offset an arbitrarily small δ-dependent piece of y into an $O(1)$ contribution to $\mu(x_*)$; see Appendix D for an explicit counterexample.*

Lemma 3 (Haar concentration)**.** *If $K = cI_n + \delta E$ is PSD with $\|E\|_{\mathrm{op}} \le 1$, $c > 0$, $0 \le \delta \le c$, and the test-train cross-covariance $\|k_*\|_\infty \le \delta'$ for $x_* \notin X$, then*

$$r_{\mathrm{eff}}(K) = n + O(n\delta/c), \quad S(K) = \log n + O(\delta/c),$$

and $\mu(x_*) = O(\delta')$, $v(x_*) = K_q(x_*,x_*) - O(\delta'^2/c)$. The posterior collapses to the prior at any point outside the training set.

Why we dropped the population asymptotic. An earlier draft of Proposition 1 additionally stated a population-level asymptotic approximating the test–train cross-covariance K_* by K^2/n (Bach, 2017, Lemma 2), giving $\mathbb{E}\,\mathrm{VC}(K) \approx 1 - \frac{1}{n}\sum_i \lambda_i^2/(\lambda_i+\sigma^2)$. We removed it: we could not establish a rigorous error bound for the approximation (what distribution "population-typical" refers to, in which norm $\mathbb{E}[k_* k_*^\top] \approx K^2/n$ holds, and why that error stays negligible after the matrix inverse and trace), and, concretely, it fails outright under our normalization. Every kernel family we study has $K_q(x,x) = 1$, so $\mathrm{tr}(K) = \sum_i \lambda_i = n$; using $\lambda_i^2/(\lambda_i+\sigma^2) = \lambda_i - \sigma^2\lambda_i/(\lambda_i+\sigma^2)$, the asymptotic simplifies, for *any* such spectrum, to $\mathbb{E}\,\mathrm{VC}(K) \approx (\sigma^2/n)\,d_\sigma(K) \in [0,\sigma^2]$ (since $0 \le d_\sigma(K) \le n$)—at Haar concentration ($K \approx I$) this evaluates to $\sigma^2/(1+\sigma^2) \approx 0.048$ for our $\sigma^2 = 0.05$, contradicting both Lemma 3 and the measured VC ≈ 0.98–0.99 at $S/\log n \approx 1$ (Figure 4). We do not use this asymptotic anywhere; VC $\to 1$ at Haar concentration follows instead

Algorithm 1 Spectral diagnosis for a Gaussian-process quantum kernel.

Require: training inputs $X = \{x_i\}_{i=1}^n$, quantum kernel $K_q(\cdot,\cdot)$ from chosen ansatz (L, s), small validation split $(X^{\mathrm{val}}, y^{\mathrm{val}})$ (optional).

$K \leftarrow [K_q(x_i, x_j)]_{i,j=1}^n$ (single Gram matrix)
Compute $S(K)/\log n$ and $r_{\mathrm{eff}}(K)/n$ from Definition 2.
if $S(K)/\log n \le 0.1$ **then**
    return regime = **constant-collapse**; recommend use only if target is known band-limited; note $r_{\mathrm{eff}}/n \lesssim 0.05$.
else
    if $S(K)/\log n \ge 0.95$ **and** measured VC ≥ 0.9 on a held-out split **then**
        return regime = **near-Haar, flagged**; posterior variance is not contracting on held-out data; consider reducing L or s before BO / active learning.
    else
        regime $\leftarrow$ **useful** (high entropy alone, *without* a corroborating high-VC/validation check, is not sufficient grounds to reject a configuration—the useful-hardness optimum itself sits near $S/\log n \approx 0.95$–0.99 for broadband targets, Sections 5, 8).
    end if
    if validation split available **then**
        Compare held-out NLL of K_q against a Matérn baseline of matched r_{eff}.
        if NLL(K_q) noticeably below baseline **then**
            return regime, *use* K_q.
        else
            return regime, *dequantize* via r_{eff}-feature classical Nyström / RFF.
        end if
    end if
    return regime; no a-priori accept/reject without validation.
end if

directly from Lemma 3's small-cross-covariance argument, and we compute VC exactly (via the finite-sample bound's true K_*) in every experiment. For completeness: the quantity $\sum_i \lambda_i^2/(\lambda_i+\sigma^2)$ that appeared in the removed asymptotic is, in any case, *not* the same as Bach's effective degrees of freedom $d_\sigma(K) := \sum_i \lambda_i/(\lambda_i+\sigma^2)$; they do not even coincide as $\sigma^2 \to 0$, tending instead to $\mathrm{tr}(K)$ and $\mathrm{rank}(K)$ respectively. Empirically, $\mathrm{VC}(K)$ ranges continuously from ≈ 0 at low spectral entropy (Lemma 2) to ≈ 1 at Haar concentration (Lemma 3).

Lemmas 2 and 3 say that both pathologies are *simultaneously* simulable (the first by the constant feature, the second by classical Gaussian random features approximating the identity) *and* useless for GP inference. The useful QGP regime must live strictly between $s(K) = 0$ and $s(K) = 1$.

E The Diagnostic Recipe

Algorithm 1 states the diagnostic procedure of Section 7 in full.

F Hardware Noise Model

Noise interpretation. The hardware bias is not random. Configurations with low simulator entropy ($S/\log n \lesssim 0.5$) are shifted *upward*; configurations near Haar concentration ($S/\log n \gtrsim 0.9$) are shifted slightly *downward* or unchanged. The following lemma decomposes the systematic part of this effect; the empirical sign in each regime arises from the combination of depolarizing and shot/readout contributions discussed in Corollary 2.

Lemma 4 (Depolarizing noise adds a rank-1 component to the Gram matrix)**.** *Let $\mathcal{E}_p(\rho) = (1-p)\rho + pI_{2^{n_q}}/2^{n_q}$ be the n_q-qubit depolarizing channel of strength $p \in [0,1]$ applied once after the compute-uncompute circuit, and let $\widehat{K}_p$ denote the kernel Gram matrix estimated under $\mathcal{E}_p$. Then for $d = 2^{n_q}$*

$$\widehat{K}_p \;=\; (1-p)\,K \;+\; \frac{p}{d}\,\mathbf{1}\mathbf{1}^\top,$$

i.e. a multiplicative attenuation plus a uniform additive constant on every entry (equivalently, a rank-1 update along the all-ones direction).

Bidirectional effect on entropy. Unlike the identity correction αI_n, the rank-1 update $(p/d)\mathbf{1}\mathbf{1}^\top$ *boosts* the eigenvalue along the all-ones direction by $\approx pn/d$ and leaves the orthogonal subspace untouched. Its effect on $s(\widehat{K}_p)$ therefore depends on the alignment of $\mathbf{1}/\sqrt{n}$ with the top eigenvectors of K: when K is Haar-like (top eigenvector roughly orthogonal to $\mathbf{1}$), the update creates a new dominant mode and *lowers* s; when K is near constant-collapse (top eigenvector already aligned with $\mathbf{1}$), the update reinforces the existing concentration and changes s very little. Realistic NISQ noise additionally contains shot, readout, and coherent components beyond the simple depolarizing model, and the observed empirical bias direction (Corollary 2) is the combined effect of all of these. We therefore treat Lemma 4 as a model for the systematic part of the hardware bias rather than a quantitative prediction of s shifts.

Corollary 2 (Empirical hardware-bias pattern)**.** *The depolarizing contribution of Lemma 4 predicts that the hardware estimate of s shifts* down *in the Haar-like regime ($s \to 1$, all-ones direction roughly orthogonal to top eigenvectors) because the rank-1 update along $\mathbf{1}$ creates a new dominant mode. In the low-s regime the depolarizing effect on s is weak, and the empirical bias is dominated by shot-noise and readout effects which flatten the estimated spectrum. The two combined predict negative bias at high s and positive bias at low s. Empirically, across our 24 XL frontier points the mean signed bias is -0.074 for $s^{\text{exact}} \geq 0.9$ and $+0.054$ for $s^{\text{exact}} \leq 0.5$, with positive Pearson correlation $\rho = 0.54$ between $(1 - s^{\text{exact}})$ and the signed bias (Appendix K). This is a noise-model observation rather than a tight theoretical prediction; (Wallman & Emerson, 2016) discusses the broader scope of NISQ noise channels beyond depolarization.*

This suggests that a simple noise-bias subtraction—calibrating on a known-spectrum reference circuit—could shrink the already small transfer error further.

G Implementation Details

G.1 Quantum ansatz families

All three families use a parameter tensor of shape $(L, n_q, 2)$ where L is the depth and n_q the qubit count. HE (*hardware-efficient*): each layer applies $R_y(s\,x_q)$ data encoding, then trainable $R_y(\theta_0)R_z(\theta_1)$, then a CNOT ring (nearest neighbor plus a closing wrap CNOT for $n_q > 2$). MATCHGATE: each layer applies trainable $R_z(\theta_0)$, then on each adjacent pair an $\mathrm{IsingXX}(s(x_q + x_{q+1})/2)$ and $\mathrm{IsingYY}(\theta_1)$. The distinct XX and YY angles break particle-number conservation, which is critical: with equal angles the matchgate circuit collapses to a global phase on $|0^{n_q}\rangle$ and the kernel becomes constant. IQP: a Hadamard sandwich enclosing $R_z(sx_q c_q^{(\ell)})$ and $\mathrm{IsingZZ}(sx_q x_{q+1} d_{q,q+1}^{(\ell)})$, where $c_q^{(\ell)} = 0.5 + 0.5\cos(\theta_0)$ and $d_{q,q+1}^{(\ell)} = 0.5 + 0.5\cos(\theta_1)$ keep the diagonal coefficients in $[0, 1]$. All trainable parameters are drawn from $\mathcal{U}[0, 2\pi)$ once per seed.

G.2 Ablation ansatze (Section 4)

RAND-PAULI-NOENT: each layer applies a single Pauli rotation per qubit, with the axis chosen by rounding θ_0 modulo 3 to $\{X, Y, Z\}$ and the angle $sx_q + \theta_1$; *no entangling gates*. CLIFFORDT: each layer applies a sparse random Clifford sublayer (per-qubit choice from $\{H, S, X, \mathbf{I}\}$ followed by a random nearest-neighbor CNOT permutation), then a data-encoding $R_z(sx_q)$ on every qubit. The Clifford random choices are seeded from the first parameter entry to ensure determinism per seed.

G.3 Classical baselines (Section 4)

We use the standard pure-Python implementations: RBF ($k(x, x') = \exp(-\|x - x'\|^2/(2h^2))$) with bandwidths $h \in \{0.1, 0.3, 0.6, 1, 1.5, 2.5, 4, 6, 10\}$; MATERN-3/2 and MATERN-5/2 at the same bandwidths; RFF: $M \in \{4, 16, 64\}$ random Fourier features at bandwidths $h \in \{0.3, 1, 3\}$, $W \sim \mathcal{N}(0, I_d)/h$, $b \sim \mathcal{U}[0, 2\pi)$; DEEP KERNEL: a two-layer `tanh` feature map of hidden width $H \in \{4, 16, 32\}$, followed by an RBF in feature space at $h \in \{0.3, 1, 3\}$; weights initialized $\mathcal{N}(0, 1/\sqrt{d_{\text{in}}})$ and $\mathcal{N}(0, 1/\sqrt{H})$, biases $\mathcal{U}[-1, 1]$, all per seed.

G.4 TARGETS

Synthetic. $y(x) = \sin(w^\top x) + 0.3\cos(2x_1) + 0.2\bar{x}$, w drawn from $\mathcal{N}(0, I_d)$ with the per-seed RNG (i.e. w varies across seeds for the multi-seed M2 and M1 sweeps).

Quantum-data. $y(x) = \langle 0^{n_q}|U_d(x)^\dagger O U_d(x)|0^{n_q}\rangle$ with $O = \frac{1}{n_q}\sum_q Z_q$ and U_d a fixed $L = 3$ data-reuploading circuit (per-qubit $R_y(sx_q + \theta_0)R_x(\theta_1)$ followed by a CNOT ring), separate from any of the kernel ansatz families. U_d parameters are drawn from $\mathcal{U}[-\pi, \pi)$ with seed 42; the seed is intentionally fixed so the quantum-data target is a single deterministic function rather than a per-seed random one (this is why the bottom row of Figure 2 carries no errorbars).

G.5 REAL-DATA BENCHMARKS

For *diabetes* (442 samples, 10 features) and *california housing* (200 random samples of 20 640, 8 features) we use the `sklearn.datasets` loaders. Features are truncated or tiled to match $n_q = 6$, standardized via `StandardScaler`, and clipped to $[-\pi, \pi]^{n_q}$; the target is centered and scaled by its sample standard deviation. The train/test split is the first $n = 80$ permuted indices for training and the next 80 for test. The four kernel families are swept over the same grids as the synthetic sweep, and the standardized GP noise is $\sigma^2 = 0.05$.

G.6 GP REGRESSION

Observation noise $\sigma^2 = 0.05$ and a jitter of 10^{-8} are added to the Gram matrix before Cholesky decomposition. The ECE uses ten equally spaced confidence levels $\{0.05, 0.15, \ldots, 0.95\}$ and a Gaussian credible interval. The variance-contraction ratio VC is the mean ratio of posterior variance to prior variance over a held-out test set of equal size.

G.7 SPECTRAL STATISTICS

S uses natural-log entropy; `r_eff` is the participation ratio $(\sum \lambda_i)^2 / \sum \lambda_i^2$; $\mathrm{dq}_{0.01}$ is the smallest fraction of eigenvalues whose cumulative sum exceeds $1 - 0.01$ of the trace; off-diagonal concentration is $\mathrm{std}(K_{\mathrm{off}})/\mathrm{mean}|K_{\mathrm{off}}|$; kernel-target alignment is $\langle K, yy^\top\rangle_F / (\|K\|_F \|yy^\top\|_F)$; centered KTA additionally projects K and $yy^\top$ through $H = I - \mathbf{1}\mathbf{1}^\top/n$; Bach's degrees of freedom uses noise $\sigma = 0.05$.

G.8 d_{int} COMPUTATION (SECTION 5)

For each (ansatz, L, s) in the M2-syn and M2-8q sweeps (Table 3) we form the training Gram matrix K, compute the eigendecomposition $K = U\Lambda U^\top$, project the training labels into the kernel basis via $\beta = U^\top y$, and report $d_{\mathrm{int}} = 1/\sum_i (\beta_i^2/\|\beta\|^2)^2$. The corollary's predicted lower bound is $s^* \geq \log d_{\mathrm{int}} / \log n$.

G.9 BO DOWNSTREAM (SECTION 8)

We run UCB Bayesian optimization with $\beta = 2.0$ on synthetic ($f = -[\sin(w^\top x) + 0.3\cos(2x_1) + 0.2\bar{x}]$) and the quantum-data ($-\langle\psi(x)|O|\psi(x)\rangle$) objectives. Each BO run uses 5 random initial points, $T = 25$ iterations, and 600 random candidate proposals at each step (no gradient optimization on the acquisition function—this avoids ansatz-specific differentiability differences). Per (surrogate, objective) cell we average over 6 seeds on the synthetic objective and 8 seeds on the quantum-data objective (Section 8). Surrogates: Q-CONSTANT (HE $L = 1$, $s = 0.10$), Q-SWEET (HE $L = 2$, $s = 0.60$), Q-HAAR (HE $L = 2$, $s = 2.00$), CLS-RBF (RBF $h = 2.5$).

G.10 HARDWARE EXPERIMENTS

M3-XL (Section 6). 24 configurations, $n_q = 4$, $n = 8$, 4096 shots, no error mitigation, `ibm_aachen` (Heron-class), Qiskit SamplerV2 primitive of `qiskit-ibm-runtime 0.47.0`, transpile `optimization_level=2`. The σ_{sim} column of Table 9 is computed from five noiseless seeds.

Resilience probe (M12, Appendix J). Same backend, single configuration (HE, $L = 3$, $s = 0.30$, $\sigma^2 = 0.05$), $n = 8$; three raw reruns intended as a `resilience_level` $\in \{0, 1, 2\}$ sweep for SamplerV2, which we discovered (see Appendix J) does not actually support that option in `qiskit-ibm-runtime 0.47`.

Cross-backend (M13, Appendix J). Six representative configurations re-run on `ibm_marrakesh` with identical settings to M3-XL.

Drift characterization (M18, Appendix J). 5 back-to-back reruns of HE $L = 3$, $s = 0.30$ on `ibm_aachen`, same X draw and ansatz parameters across reruns.

Six-qubit scale-up (M19, Appendix J). 9 balanced configurations ($n_q = 6$, $n = 8$) on `ibm_aachen`; the cross-Gram matrix contains 36 pairwise circuits per configuration (324 circuits total).

G.11 Seed protocol

Unless otherwise noted, the per-seed RNG draws (i) training inputs X_{tr} and (ii) all trainable kernel parameters θ; observation noise and the quantum-data target circuit U_d are deterministic (fixed seeds 0 and 42 respectively) so that the quantum-data observable is a single deterministic function across seeds. The multi-seed sweeps reported in Figures 4, 2 (top row), and Section 8 average over 3–10 seeds as specified.

G.12 Compute

All simulations were run on commodity GPUs (one NVIDIA RTX A6000 per sweep). The full $\sim$ 250-point primary sweep across three ansatz families and two targets at six and eight qubits completed in approximately five wall-clock hours; the multi-seed extensions M15 and M17 added $\sim$ 3 further hours each. Across the *entire* hardware campaign—M3-XL (24 pts, `ibm_aachen`), the M13 cross-backend check (6 pts, `ibm_marrakesh`), the M18 drift reruns (5 pts, `ibm_aachen`), and the M19 six-qubit scale-up (9 pts, `ibm_aachen`)—total IBM backend usage time across both devices was approximately 80 minutes; see Appendix I for the M3-XL sweep's own execution time in isolation.

H Full Proofs

Proof of Lemma 1

The first inequality holds because $r_{\text{eff}}(K) = 1/\sum_i \tilde{\lambda}_i^2$ and $\sum_i \tilde{\lambda}_i^2 \le 1$ (with equality iff one $\tilde{\lambda}_i$ equals 1). The Rényi-2 entropy of $\{\tilde{\lambda}_i\}$ is $H_2 = -\log \sum_i \tilde{\lambda}_i^2 = \log r_{\text{eff}}(K)$. Since the Shannon entropy $S = H_1 \ge H_2$ for any discrete probability distribution (standard ordering of Rényi entropies), $\log r_{\text{eff}}(K) \le S(K)$, i.e. $r_{\text{eff}} \le e^S$. Finally, $S/\log n = s$ gives $e^S = n^s$, and $s \le 1$ implies $r_{\text{eff}} \le n^s \le n$. □

Proof of Lemma 2

Write $K = c\mathbf{1}\mathbf{1}^\top + \delta E$. The rank-1 base has eigenvalues $(nc, 0, \ldots, 0)$, so $r_{\text{eff}} = 1$ and $S = 0$. By Weyl's inequality, $|\lambda_i(K) - \lambda_i(c\mathbf{1}\mathbf{1}^\top)| \le \delta\|E\|_{\text{op}} \le \delta$ for every i, and PSD-ness of K gives $\lambda_1(K) \in [nc - \delta, nc + \delta]$, $\lambda_i(K) \in [0, \delta]$ for $i \ge 2$. The participation ratio $r_{\text{eff}} = (\sum \lambda)^2/\sum \lambda^2$ is a smooth function of the eigenvalues away from degeneracies, with first-order perturbation $\delta/(nc) \cdot O(1)$, giving $r_{\text{eff}}(K) = 1 + O(\delta/c)$.

The Shannon entropy is *not* similarly Lipschitz at this rank-1 spectrum: unlike r_{eff}, $-x \log x$ has unbounded derivative as $x \to 0^+$, so a δ-sized perturbation can move the entropy by more than $O(\delta)$. Let $\tilde{\lambda}_i := \lambda_i(K)/\text{tr}(K)$ with $\text{tr}(K) = nc + O(\delta) = nc(1 + O(\delta/c))$, and let $\varepsilon := 1 - \tilde{\lambda}_1 = \sum_{i \ge 2} \tilde{\lambda}_i \le (n-1)\delta/\text{tr}(K) = O(\delta/c)$. Among distributions on $\{1, \ldots, n\}$ with $\tilde{\lambda}_1 = 1 - \varepsilon$ fixed,

entropy is maximized by spreading the remaining mass ε uniformly over the other $n-1$ coordinates, so

$$S(K) \;\le\; -(1-\varepsilon)\log(1-\varepsilon) \;-\; \varepsilon \log \frac{\varepsilon}{n-1} \;=\; O(\varepsilon) \;+\; \varepsilon \log(n-1) \;+\; \varepsilon \log(1/\varepsilon).$$

For fixed n, as $\varepsilon = O(\delta/c) \to 0$ the $\varepsilon \log(1/\varepsilon)$ term dominates the two $O(\varepsilon)$ terms, giving $S(K) = O(\varepsilon \log(1/\varepsilon)) = O\big((\delta/c)\log(c/\delta)\big)$. For the posterior with $\sigma^2 > 0$ *fixed*, write $k_* = c\mathbf{1} + \delta' e$ per the stated hypothesis. First take $\delta' = 0$ exactly ($k_* = c\mathbf{1}$): $K + \sigma^2 I = c\mathbf{1}\mathbf{1}^\top + \sigma^2 I + \delta E$; applying the Sherman–Morrison identity, $(c\mathbf{1}\mathbf{1}^\top + \sigma^2 I)^{-1} y = \sigma^{-2} y - \frac{c}{\sigma^4 + nc\sigma^2}\mathbf{1}(\mathbf{1}^\top y)$, and using $(c\mathbf{1}\mathbf{1}^\top + \sigma^2 I)^{-1}\mathbf{1} = \frac{1}{\sigma^2 + nc}\mathbf{1}$ (same identity applied to $y = \mathbf{1}$),

$$\mu(x_*) = k_*^\top (K + \sigma^2 I)^{-1} y = c\,\mathbf{1}^\top (c\mathbf{1}\mathbf{1}^\top + \sigma^2 I)^{-1} y + O_\sigma(\delta/c) = \frac{nc}{nc + \sigma^2}\,\bar{y} + O_\sigma(\delta/c),$$

$$v(x_*) = K_q(x_*, x_*) - k_*^\top (K + \sigma^2 I)^{-1} k_* = K_q(x_*, x_*) - c^2\,\mathbf{1}^\top (c\mathbf{1}\mathbf{1}^\top + \sigma^2 I)^{-1}\mathbf{1} + O_\sigma(\delta/c) = K_q(x_*, x_*) - \frac{nc^2}{nc + \sigma^2} + O_\sigma(\delta/c)$$

where O_σ hides a constant that grows as $\sigma^2 \to 0$ (the perturbation term involves $(c\mathbf{1}\mathbf{1}^\top + \sigma^2 I)^{-1}\delta E$, whose norm is controlled by δ/σ^2, not δ alone); this is why the lemma requires σ^2 fixed away from 0, and does *not* claim $\mu(x_*) \to \bar{y}$ (only to the *shrunk* mean $\frac{nc}{nc+\sigma^2}\bar{y}$) even as $\delta \to 0$ with σ^2 fixed, let alone if $\sigma^2 \to 0$ together with $\delta \to 0$. For $\delta' \neq 0$, writing $k_* = c\mathbf{1} + \delta' e$ and expanding $\mu(x_*)$, $v(x_*)$ to first order in δ' contributes an additional $O_\sigma(\delta'/c)$ term by the same Sherman–Morrison bound applied to the $\delta' e$ piece, giving the stated $O_\sigma(\max(\delta, \delta')/c)$.

The exactly noiseless case can fail. Take $n = 2$, $X = \{-1, 1\}$, $y = (-1, 1)$, and the single-qubit fidelity kernel $k_t(x, x') = \cos^2(t(x - x'))$ for $t \in (0, \pi/4)$, a valid PSD kernel (Bochner). Then $K = \cos^2(2t)\,\mathbf{1}\mathbf{1}^\top + \sin^2(2t)\,I_2$ exactly, matching the lemma's hypothesis with $c = \cos^2(2t) \to 1$ and $\delta = \sin^2(2t) \to 0$ as $t \to 0$. The noiseless ($\sigma^2 = 0$) prediction at $x_* = 1/2$ is

$$\mu(x_*) = k_*^\top K^{-1} y = \frac{\sin(2t)\sin(t)}{\sin^2(2t)} = \frac{1}{2\cos(t)} \;\longrightarrow\; \frac{1}{2} \quad \text{as } t \to 0,$$

whereas $\bar{y} = 0$: the noiseless posterior mean does *not* converge to $\bar{y}$ even though $\delta \to 0$. The small eigenvalue $\sin^2(2t)$ of K is inverted exactly, amplifying the component of y along that direction without bound as $t \to 0$; a fixed $\sigma^2 > 0$ regularizes this and restores $\mu(x_*) \to \bar{y}$. □

Proof of Lemma 3

Analogous to Lemma 2 with $K = cI + \delta E$ giving eigenvalues $(c, \ldots, c)$ perturbed by $O(\delta)$, hence $r_{\text{eff}} = n + O(n\delta/c)$ and $S = \log n - O(\delta/c)^2$. The posterior mean $k_*^\top (K + \sigma^2 I)^{-1} y = O(\|k_*\|_\infty \cdot \|y\|/(c + \sigma^2))$ is $O(\delta')$ when $\|k_*\|_\infty \le \delta'$. □

Proof of Theorem 1

Sort $\tilde{\lambda}_i$ in decreasing order. By Cauchy–Schwarz applied to the indicator $\mathbf{1}_{i>k}$ and the vector $(\tilde{\lambda}_i)$,

$$\Big(\sum_{i>k} \tilde{\lambda}_i\Big)^2 \;\le\; (n-k) \cdot \sum_{i>k} \tilde{\lambda}_i^2 \;\le\; (n-k) \cdot \sum_{i=1}^{n} \tilde{\lambda}_i^2 \;=\; \frac{n-k}{r_{\text{eff}}(K)}.$$

Hence $\sum_{i>k} \tilde{\lambda}_i \le \sqrt{(n-k)/r_{\text{eff}}(K)}$, and the equivalence between $\sum_{i>k} \tilde{\lambda}_i$ and the nuclear-norm tail $\|K - K_k\|_*/\|K\|_*$ is Eckart–Young.

Substituting the high-rank assumption $r_{\text{eff}}(K) \ge \rho\, n$ gives $\sum_{i>k} \tilde{\lambda}_i \le \sqrt{(n-k)/(\rho n)}$, and solving $\sqrt{(n-k)/(\rho n)} = \varepsilon$ yields the rank-deficiency bound $n - k \le \rho\, n\, \varepsilon^2$.

Why no entropy-based extension. The natural temptation is to chain the Cauchy–Schwarz bound with Lemma 1 ($r_{\text{eff}} \le n^{s(K)}$) to obtain $\sqrt{(n-k)/r_{\text{eff}}} \le \sqrt{(n-k)n^{-s(K)}}$. This step has the wrong direction: $r_{\text{eff}} \le n^s$ implies $1/r_{\text{eff}} \ge n^{-s}$, so the inequality flips, and no useful bound on the tail follows from the spectral entropy alone. We therefore restrict the statement to the high-rank regime $r_{\text{eff}} \ge \rho n$, which corresponds to near-Haar spectra and is precisely the regime where Nyström compressibility is informative. □

PROOF OF PROPOSITION 1

Posterior variance: $v(x_*) = K_q(x_*, x_*) - k_*^\top (K + \sigma^2 I)^{-1} k_*$. Averaging over the test set,

$$\frac{1}{n_*} \sum_{j=1}^{n_*} v(x_*^{(j)}) \;=\; \overline{K_{q_*}} \;-\; \frac{1}{n_*} \operatorname{tr}\Big(K_*^\top (K + \sigma^2 I)^{-1} K_* \Big),$$

which gives the stated finite-sample lower bound after dividing by $\min_j K_q(x_*^{(j)}, x_*^{(j)})$; this holds for the true K_* with no further assumptions. □

(An earlier version of this proof additionally attempted an in-expectation reduction via $\mathbb{E}_{x_*}[k_* k_*^\top] \approx K^2/n$, following Bach, 2017, Lemma 2, to relate $\mathbb{E}\,\mathrm{VC}(K)$ to $\sum_i \lambda_i^2/(\lambda_i + \sigma^2)$ and hence to Bach's degrees of freedom $d_\sigma(K)$ as $\sigma^2 \to 0$. We removed this step throughout the paper (Section 3, Appendix D): we could not justify the approximation rigorously, and, concretely, it evaluates to $\lesssim \sigma^2$ under our kernel normalization for *every* spectrum, contradicting the measured VC ≈ 0.98 at Haar concentration. Only the finite-sample bound above is used or claimed anywhere in this paper.)

PROOF OF THEOREM 2

Fix the eigenbasis $\{u_i\}$, write $K = \sum_i \lambda_i u_i u_i^\top$ with $\lambda_i \geq 0$, and let $\beta_i := u_i^\top y$. Since $K + \sigma^2 I = \sum_i (\lambda_i + \sigma^2) u_i u_i^\top$, the quadratic form decomposes as

$$y^\top (K + \sigma^2 I)^{-1} y \;=\; \sum_i \frac{\beta_i^2}{\lambda_i + \sigma^2}, \qquad \log\det(K + \sigma^2 I) \;=\; \sum_i \log(\lambda_i + \sigma^2),$$

and the negative marginal log likelihood becomes the separable sum $\mathcal{L}(\lambda) = \frac{1}{2} \sum_i \mathcal{L}_i(\lambda_i) + \text{const}$ with $\mathcal{L}_i(\lambda) := \beta_i^2/(\lambda + \sigma^2) + \log(\lambda + \sigma^2)$.

Unconstrained minimum. $\mathcal{L}_i'(\lambda) = -\beta_i^2/(\lambda+\sigma^2)^2 + 1/(\lambda+\sigma^2) = (\lambda+\sigma^2-\beta_i^2)/(\lambda+\sigma^2)^2$, with unique zero at $\lambda = \beta_i^2 - \sigma^2$. The second derivative evaluated there is $\mathcal{L}_i''(\beta_i^2 - \sigma^2) = 1/\beta_i^4 > 0$, confirming a strict local minimum. For $\beta_i^2 < \sigma^2$ the unique critical point is in the infeasible region $\lambda < 0$: since $\lambda \geq 0$ forces $\lambda + \sigma^2 \geq \sigma^2 > \beta_i^2$, the numerator $\lambda + \sigma^2 - \beta_i^2$ stays strictly positive throughout the feasible region, so $\mathcal{L}_i$ is *increasing* on $\lambda \geq 0$; hence the constrained minimum is at the boundary $\lambda_i^* = 0$. Combining, $\lambda_i^* = \max(0, \beta_i^2 - \sigma^2)$.

Trace-budget minimum (implicit form). Introduce the Lagrangian $\mathcal{L}(\lambda) - (-\eta)\big(\sum_i \lambda_i - T\big)$ with $\eta \geq 0$ the dual variable for the trace inequality (the sign is chosen so that $\eta = -\partial\mathcal{L}/\partial T \geq 0$). KKT stationarity for $\lambda_i > 0$ gives $\mathcal{L}_i'(\lambda_i) + \eta = 0$, i.e., $(t - \beta_i^2)/t^2 + \eta = 0$ with $t = \lambda_i + \sigma^2$, which rearranges to

$$\eta\, t^2 + t - \beta_i^2 = 0.$$

The discriminant $1 + 4\eta\beta_i^2 \geq 0$ is always non-negative, and the unique positive root is

$$t_i^*(\eta) \;=\; \frac{-1 + \sqrt{1 + 4\eta\beta_i^2}}{2\eta}.$$

Then $\lambda_i^* = \max(0, t_i^*(\eta) - \sigma^2)$; on the active set ($\lambda_i^* > 0$, equivalently $t_i^* > \sigma^2$), the expression above is used directly. The dual variable η is determined by the constraint $\sum_i \lambda_i^*(\eta) = T$, which has a unique solution since each $t_i^*(\eta)$ is strictly decreasing in η on $\eta > 0$ (so the sum is strictly decreasing in η).

High-SNR expansion. For $\eta\beta_i^2$ small, expand $\sqrt{1 + 4\eta\beta_i^2} = 1 + 2\eta\beta_i^2 - 2\eta^2\beta_i^4 + O(\eta^3\beta_i^6)$, giving

$$t_i^*(\eta) \;=\; \beta_i^2 \;-\; \eta\,\beta_i^4 \;+\; O(\eta^2\beta_i^6),$$

so $\lambda_i^* \;\approx\; \beta_i^2 - \eta\beta_i^4 - \sigma^2$. The correction is β_i-dependent (proportional to β_i^4), *not* a uniform multiplicative rescaling $\alpha\beta_i^2$. Recovering the unconstrained MLE in the limit $\eta \to 0$ requires the β_i^4 correction term to vanish.

Effective rank. In the high-SNR regime ($\eta\beta_i^2 \ll 1$, $\beta_i^2 \gg \sigma^2$ on the active set), $\lambda_i^* \approx \beta_i^2$ and

$$r_{\mathrm{eff}}(K^*) = \frac{(\sum_i \lambda_i^*)^2}{\sum_i (\lambda_i^*)^2} \;\approx\; \frac{(\sum_i \beta_i^2)^2}{\sum_i \beta_i^4} \;=\; \frac{1}{\sum_i \hat{\beta}_i^2} \;=\; d_{\mathrm{int}}(y, K^*),$$

the inverse participation ratio of the target coordinates. Outside the high-SNR limit, $r_{\text{eff}}(K^*) \approx d_{\text{int}}$ with corrections of order $\eta\beta^2_{\max}$ from the dual variable and $\sigma^2/\beta^2_{\min}$ from the regularizer on the active set. □

PROOF OF LEMMA 4

Let $\rho_{x,x'} = U_\phi(x')^\dagger U_\phi(x)\,|0\rangle\langle 0|\,U_\phi(x)^\dagger U_\phi(x')$ denote the pre-measurement state of the noiseless compute-uncompute circuit; the noiseless kernel is $K_q(x,x') = \langle 0|\rho_{x,x'}|0\rangle$. Under a single depolarizing channel $\mathcal{E}_p$ applied at the end of the circuit, $\widehat{\rho}_{x,x'} = (1-p)\rho_{x,x'} + p\,I/d$ with $d = 2^{n_q}$, so

$$\widehat{K_q}(x,x') \;=\; \langle 0|\widehat{\rho}_{x,x'}|0\rangle \;=\; (1-p)\,K_q(x,x') \;+\; \frac{p}{d}.$$

The additive constant p/d is the same for every (x,x') pair, so stacking into a Gram matrix gives $\widehat{K}_p = (1-p)\,K + (p/d)\,\mathbf{1}\mathbf{1}^\top$ *exactly*, with no $O(\cdot)$ correction. The update is rank-1 along the all-ones direction $\mathbf{1}/\sqrt{n}$. Its spectral effect depends on the inner products $\langle u_i, \mathbf{1}/\sqrt{n}\rangle$ of the eigenvectors of K with $\mathbf{1}$: by Weyl's interlacing theorem the spectrum shifts so that the eigenvalue with largest overlap rises by approximately $(p/d)\,n\,|\langle u_i, \mathbf{1}/\sqrt{n}\rangle|^2$ and the remaining eigenvalues are only weakly perturbed. Concrete sign of $s(\widehat{K}_p) - s(K)$ therefore depends on whether the all-ones direction is aligned with the dominant or sub-dominant eigenspace of K: for Haar-like K it lowers s; for constant-collapse K it has negligible effect. The empirical bias in the low-s regime is therefore dominated by shot and readout noise rather than by depolarization, and Lemma 4 should be read as the systematic depolarizing contribution only. □

I EXPERIMENTAL CONFIGURATIONS

Table 3 summarizes every sweep that produced the body figures. Random seeds, jitter, GP noise, and ECE binning are identical across sweeps (Appendix G). The IBM-hardware sweep is reported separately in Table 4.

Table 3: Simulator sweeps. Each entry of the grid is a single (ansatz, L, s) kernel evaluation: n^2 pairwise kernel calls plus one n_*-shot test-Gram evaluation, GP posterior, and four spectral statistics. Total sweep count is |ansatze| · |depths| · |scales|.

Sweep	n_q	n / n_*	depths L	scales s	ansatze	target
M1 (Fig. 4)	6	30/30	$\{1,2,3,5,8,12\}$	$\{0.1, 0.5, 1, 2\}$	he	synthetic
M2-syn (Fig. 2 top)	6	30/30	$\{1,2,3,5,8\}$	$\{0.1, 0.3, 0.6, 1, 1.5\}$	he/mg/iqp	synthetic
M2-qt (Fig. 2 bot.)	6	30/30	$\{1,2,3,5,8\}$	$\{0.1, 0.3, 0.6, 1, 1.5\}$	he/mg/iqp	quantum
M2-8q (Fig. 9)	8	40/40	$\{1,2,3,4,6,8\}$	$\{0.1, 0.3, 0.6, 1, 1.5\}$	he/mg/iqp	synthetic

Table 4: IBM-hardware sweep (Fig. 10, Table 1). All circuits are executed on `ibm_aachen` (*Heron*-class, 156 qubits) via the Qiskit SamplerV2 primitive of `qiskit-ibm-runtime 0.47.0`, with default optimization level 2 and no error mitigation. The cross-backend probe of Appendix J re-runs 6 configurations on `ibm_marrakesh`.

Backend (main)	`ibm_aachen` (156-qubit Heron)
Cross-backend probe	`ibm_marrakesh` (156-qubit Heron)
Plan	internal premium-tier access
Used qubits	4
Training set size	$n = 8$ uniformly in $[-\pi, \pi]^4$
Shots per circuit	4096
Circuits / point	$n(n+1)/2 = 36$ (compute-uncompute pairs)
Total circuits (XL sweep)	$24 \times 36 = 864$
Wall-clock (queue + exec)	≈38 min for all 24 points
Error mitigation	none (see Appendix J for a failed attempt at a resilience-level probe)
Frontier coverage	$s \in [0.19, 1.00]$, 3 ansatze (he, mg, iqp)

Software stack. All simulator results were produced with `PennyLane 0.42.3`, `lightning.qubit` backend where available, `numpy 2.2.6`, `scipy 1.15.3`. Hardware results used `qiskit 2.4.1` and `qiskit-ibm-runtime 0.47.0`. All code is available in the supplementary material.

Compute. Simulator sweeps were run on a single NVIDIA RTX A6000 GPU; total wall-clock ≈ 5 hours for all four sweeps. The M3-XL sweep of Table 4 alone—the 24-point, `ibm_aachen` sweep behind Table 1 and Figure 10—cost approximately 9 minutes of IBM backend *execution* time (QPU-busy time, excluding queue wait), versus the ≈ 38 minute queue-plus-execution wall-clock reported in Table 4; the ≈ 80 minute figure of Appendix G is the backend-usage total across the entire hardware campaign (M3-XL plus the cross-backend, drift, and 6-qubit-scale-up sweeps of Appendix J, split across `ibm_aachen` and `ibm_marrakesh`).

J Cross-Backend and Resilience-Mitigation Validation

We probe two further dimensions of hardware robustness beyond the 24-point sweep of Table 1.

Cross-backend transfer. We pick six representative configurations spanning the frontier and re-run them on a *different* 156-qubit IBM Heron device, `ibm_marrakesh`, with identical settings ($n_q = 4$, $n = 8$, 4096 shots, no error mitigation). The diagnostic transfers with mean $|\Delta S/\log n| = 0.027$ and worst case 0.044 (Table 5), comparable to (and on the depth-stratified subset better than) the `ibm_aachen` numbers. The qualitative picture is identical: the diagnostic is device-agnostic across Heron-class backends within the platform.

Table 5: Cross-backend validation: six configurations re-run on `ibm_marrakesh`.

ansatz	L	s	exact $S/\log n$	`ibm_marrakesh`	$\lvert\Delta\rvert$
he	2	0.20	0.628	0.608	0.021
he	2	0.50	0.949	0.946	0.003
matchgate	2	0.30	0.277	0.319	0.042
matchgate	2	0.50	0.392	0.432	0.040
iqp	1	0.20	0.553	0.596	0.044
iqp	2	0.30	0.861	0.873	0.012
mean / max					0.027 / 0.044

Hardware errors are time-correlated. We re-ran the worst HE configuration from Table 9 (HE, $L = 3$, $s = 0.30$, where the original `ibm_aachen` run returned $|\Delta S/\log n| = 0.300$) three times at a later time on the same backend. We had intended this as a sweep of the SamplerV2 `resilience_level` option ($\in \{0, 1, 2\}$, i.e. raw / readout mitigation / TREX); on auditing the run script for this submission we found that `qiskit-ibm-runtime 0.47`'s `SamplerOptions` does not expose a `resilience_level` field (unlike `EstimatorV2`), so the assignment raised a validation error that a broad `try/except` in our code silently swallowed. All three reruns therefore used the identical default (unmitigated) sampler configuration; we report them as three independent raw reruns rather than a resilience sweep. They gave $|\Delta| = 0.005, 0.008, 0.012$ respectively—25–60$\times$ smaller than the original—which is consistent with the 0.300 outlier being a transient calibration/drift event of the physical device rather than a reproducible property of this configuration, though a single day's worth of reruns cannot fully rule out other explanations; we can no longer claim anything about the effect of resilience options themselves.

The combined picture is that the spectral diagnostic transfers across backends to within a few percent of the noiseless value in typical conditions, with occasional larger outliers (presumably tied to local calibration drift) that disappear on rerun; we were not able to test resilience settings specifically (see above).

Drift characterization: 5 back-to-back reruns of the worst configuration. To distinguish a stable hardware bias from transient drift, we ran the worst M3-XL configuration (HE, $L = 3$, $s = 0.30$, original $|\Delta S| = 0.300$) five additional times back-to-back on `ibm_aachen` (Table 7).

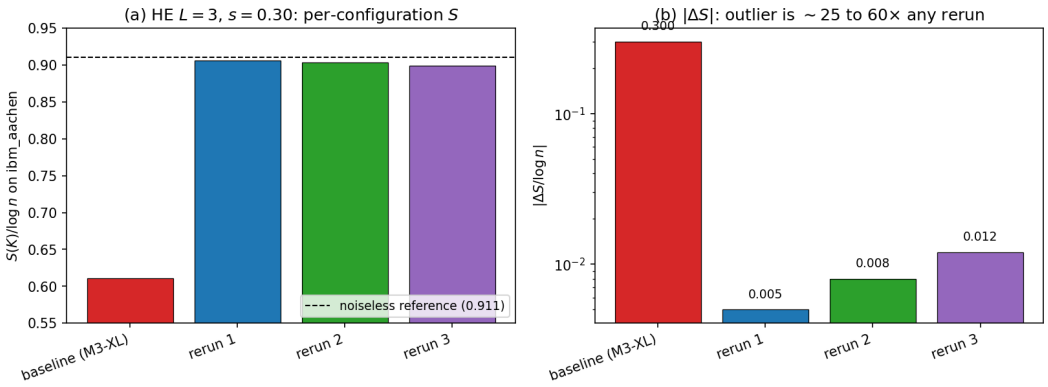


Figure 12: Three raw reruns (no mitigation; see text) of the previously-worst HE configuration, relabeled "rerun 1/2/3" (the original plot's "resilience level 0/1/2" labels were an artifact of the same script bug and have been corrected). (a) Hardware estimate of $S/\log n$ on each rerun; all three sit within 0.012 of the noiseless reference (0.911), while the original M3-XL run returned 0.611. (b) $|\Delta S/\log n|$: log-scale comparison shows the original outlier is 25–60$\times$ any individual rerun's deviation, consistent with (though not conclusive proof of) a transient calibration event rather than a reproducible property of this configuration.

Table 6: Three raw reruns of the previously-worst HE configuration on a later day, same backend, identical (unmitigated) sampler settings; see text for why this is not a resilience-level sweep. $\Delta S/\log n$ is hardware $-$ noiseless simulator.

setting	$S/\log n$	$\lvert\Delta S/\log n\rvert$	comment
noiseless simulator	0.911	–	reference
rerun 1	0.906	0.005	raw, unmitigated
rerun 2	0.904	0.008	raw, unmitigated
rerun 3	0.899	0.012	raw, unmitigated
original M3-XL run (different day)	–	0.300	transient drift

All five reruns return $|\Delta S| \in [0.001, 0.005]$ with mean 0.003 and standard deviation 0.001, i.e. within the noiseless five-seed sampling variation of $\sigma_{\text{sim}} = 0.042$ reported in Table 9. The original 0.300 outlier is therefore not reproducible: five independent reruns of the identical configuration all land within 0.5% of the noiseless reference, consistent with (though, as with the M12 reruns above, not conclusive proof of) a transient calibration event rather than a stable property of this configuration or of the diagnostic.

Table 7: Drift characterization. The worst M3-XL HE configuration (HE, $L = 3$, $s = 0.30$, $S^{\text{exact}} = 0.911$) re-run five consecutive times on `ibm_aachen`; mean $|\Delta S| = 0.003$, max 0.005, std 0.001.

rerun	S^{ibm}	$\lvert\Delta S\rvert$
1	0.909	0.003
2	0.906	0.005
3	0.909	0.002
4	0.908	0.003
5	0.910	0.001
original M3-XL run	0.611	0.300

Combined hardware view. Figure 14 overlays all four hardware sweeps—M3-XL ($n_q = 4$ on `ibm_aachen`, 24 pts), M13 cross-backend (`ibm_marrakesh`, 6 pts), M18 drift (5 reruns of the worst HE config), and the M19 6-qubit scale-up (9 pts)—on a single simulator-vs.-hardware scatter.

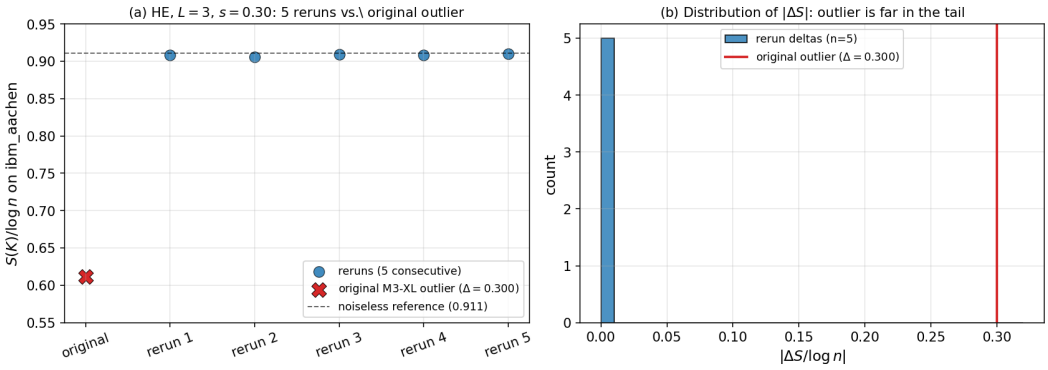


Figure 13: Hardware drift: five consecutive reruns of the worst M3-XL configuration on `ibm_aachen` cluster tightly around the noiseless reference (dashed line), while the original outlier (red X) sits 0.30 below. (a) Per-rerun scatter: the five reruns agree on $S/\log n \in [0.906, 0.910]$, indistinguishable from the noiseless 0.911 within sampling variance. (b) Histogram of $|\Delta S/\log n|$: all five reruns are at ≤ 0.005; the original outlier at 0.300 is far in the tail.

Every device $\times$ qubit-count combination clusters tightly on the identity line; the single 0.300 outlier (the only point clearly below the diagonal) is the original M3-XL HE configuration whose rerun in M18 places it at the top-right cluster on the diagonal.

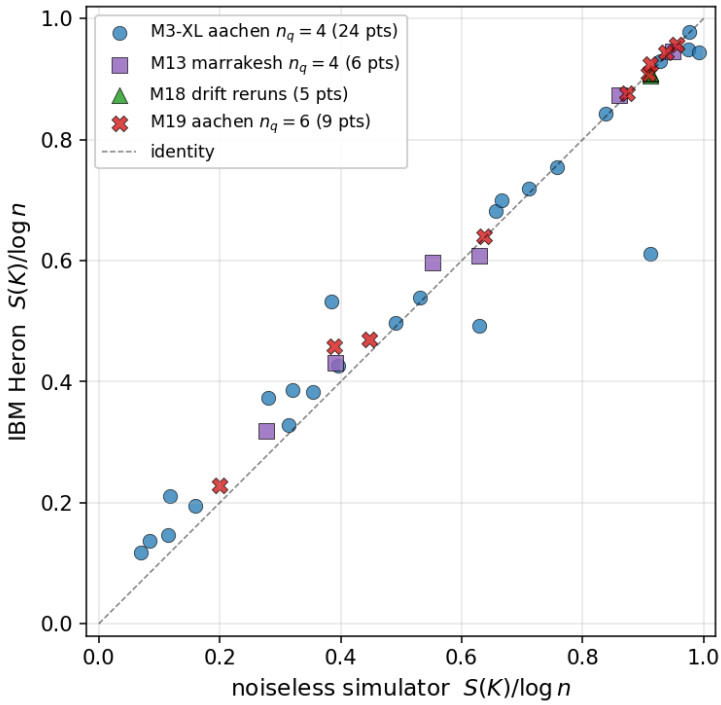


Figure 14: Hardware overview. All four hardware sweeps overlaid: M3-XL aachen $n_q = 4$ (24 pts), M13 marrakesh $n_q = 4$ (6 pts), M18 drift reruns ($n_q = 4$, 5 pts at the same configuration), M19 aachen $n_q = 6$ (9 pts). All sweeps cluster on the identity line; the single visible outlier at $(S^{\text{exact}}, S^{\text{ibm}}) \approx (0.91, 0.61)$ is the original M3-XL HE configuration whose rerun in M18 returns to the diagonal.

Six-qubit scaling probe. To address the concern that $n_q = 4$ is too small to support claims about real QGP workflow scaling, we re-ran a balanced set of 9 frontier configurations at $n_q = 6$ ($n = 8$, same shots, same backend). The diagnostic transfers with mean $|\Delta S| = 0.017$ and max 0.069, which is *better* than the corresponding $n_q = 4$ numbers (Table 8). Per family, HE and matchgate transfer within ≤ 0.028 worst-case error; IQP shows the largest single deviation (0.069) but mean $\leq 3\%$. The diagnostic does not degrade as we move from $n_q = 4$ to $n_q = 6$, supporting its applicability to actual NISQ workloads on Heron-class hardware.

Table 8: Hardware scale-up to $n_q = 6$ on `ibm_aachen`, 9 balanced configurations.

#	ansatz	L	s	S^{exact}	S^{ibm}	$\lvert\Delta S\rvert$
1	he	1	0.10	0.200	0.229	0.028
2	he	1	0.50	0.955	0.957	0.002
3	he	2	0.30	0.909	0.908	0.001
4	matchgate	1	0.30	0.447	0.470	0.023
5	matchgate	2	0.30	0.636	0.641	0.004
6	matchgate	2	0.50	0.874	0.877	0.003
7	iqp	1	0.10	0.390	0.459	0.069
8	iqp	1	0.30	0.912	0.924	0.012
9	iqp	2	0.20	0.937	0.943	0.006
mean / max						0.017 / 0.069

K Hardware Configuration Detail

Table 9 reports the full per-configuration data for the 24 hardware points of Section 6. σ_{sim} is the standard deviation of $S/\log n$ across five random seeds of the training inputs and ansatz parameters at the same (ansatz, L, s) on the noiseless statevector simulator. We use it as a coarse normalization for $z = |\Delta S|/\sigma_{\text{sim}}$, but note it measures configuration-resampling spread, not hardware or shot noise, so z is a rough same-order-of-magnitude check rather than a formal significance test; see Appendix J for a direct measurement of hardware run-to-run variability on a fixed configuration.

Table 9: Full per-configuration hardware data (`ibm_aachen`, $n_q = 4$, $n = 8$, 4096 shots, no error mitigation). HE shows the largest hardware-simulator discrepancy, especially at depth $L \geq 3$; matchgate and IQP transfer with $z < 2$ on most points.

#	ansatz	L	s	S^{exact}	S^{aer}	S^{ibm}	σ_{sim}	z
1	he	1	0.10	0.160	0.162	0.195	0.013	2.6
2	he	1	0.30	0.656	0.654	0.682	0.043	0.6
3	he	2	0.10	0.281	0.285	0.373	0.039	2.4
4	he	2	0.20	0.628	0.626	0.492	0.067	2.0
5	he	2	0.50	0.975	0.975	0.950	0.028	0.9
6	he	2	0.80	0.992	0.992	0.944	0.008	5.8
7	he	3	0.10	0.385	0.385	0.533	0.039	3.8
8	he	3	0.30	0.911	0.911	0.611	0.042	7.2
9	matchgate	1	0.10	0.083	0.087	0.136	0.014	3.7
10	matchgate	1	0.30	0.396	0.400	0.427	0.056	0.5
11	matchgate	2	0.10	0.070	0.073	0.117	0.018	2.5
12	matchgate	2	0.30	0.314	0.312	0.328	0.066	0.2
13	matchgate	2	0.50	0.531	0.531	0.539	0.077	0.1
14	matchgate	2	0.80	0.758	0.757	0.755	0.054	0.1
15	matchgate	3	0.10	0.114	0.117	0.147	0.022	1.5
16	matchgate	3	0.30	0.490	0.490	0.498	0.066	0.1
17	iqp	1	0.05	0.118	0.119	0.211	0.030	3.1
18	iqp	1	0.10	0.320	0.319	0.386	0.075	0.9
19	iqp	1	0.20	0.666	0.666	0.700	0.129	0.3
20	iqp	1	0.30	0.838	0.837	0.843	0.125	0.0
21	iqp	2	0.05	0.354	0.353	0.383	0.098	0.3
22	iqp	2	0.10	0.710	0.709	0.720	0.167	0.1
23	iqp	2	0.20	0.928	0.927	0.930	0.132	0.0
24	iqp	2	0.30	0.977	0.977	0.978	0.083	0.0

L SCALING TO $n = 100$

To address concerns about the modest training-set size of the main experiments ($n = 30$ at $n_q = 6$ and $n = 40$ at $n_q = 8$), we re-ran the M1 sweep (HE family, $n_q = 6$, depths $L \in \{1, 2, 3, 5, 8, 12\}$, scales $s \in \{0.1, 0.5, 1, 2\}$) at $n = 100$. The full sweep took ≈3.3 hours on a single A6000 with `lightning.qubit`. The qualitative trends are unchanged but two quantitative shifts are worth noting (Figure 15):

1. The **NLL-optimal configuration** $(L, s) = (1, 0.5)$ **is invariant** between $n = 30$ and $n = 100$.
2. The **best NLL improves** from $+0.89$ at $n = 30$ to $+0.46$ at $n = 100$, as expected from more training data.
3. At fixed (L, s) the normalized spectral entropy $S(K)/\log n$ **decreases** when n grows (right panel), because $\log n$ grows faster than the spectral entropy itself. Consequently the absolute sweet-spot $S(K^*)/\log n$ shifts from 0.91 at $n = 30$ to 0.79 at $n = 100$.

Combined with the n_q-dependent shift discussed in Section 5 (0.91 at $n_q = 6$ versus 0.99 at $n_q = 8$), this is consistent with comparing against Corollary 1's interval-valued prediction rather than a fixed absolute threshold, since the sweet-spot value depends on both the training-set size n and the qubit count n_q; as noted in Section 5, this consistency is descriptive, not a validated confirmation of the corollary's mechanism, given the idealized-optimization mismatch discussed there.

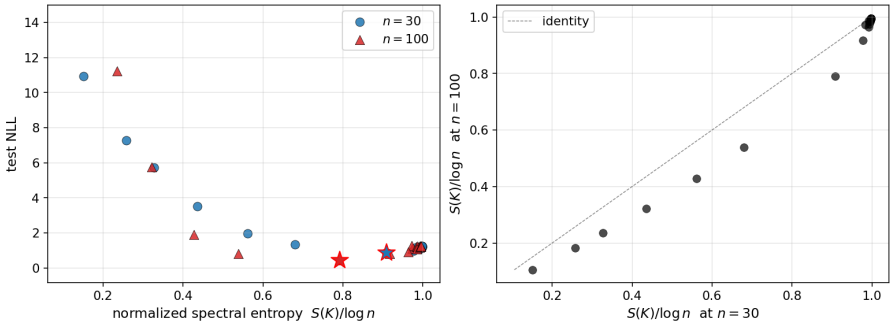


Figure 15: Scaleup from $n = 30$ to $n = 100$ on the synthetic target, $n_q = 6$, HE family. (Left) The same NLL U-shape; stars mark the best-NLL configuration in each sweep, which is identical in (L, s) space but lies at a different $S(K)/\log n$ value (0.79 vs. 0.91) because $\log n$ grew. (Right) Per-configuration $S(K)/\log n$ at $n = 100$ vs. $n = 30$; all points lie below the identity line, confirming the $\log n$ normalization effect.

M SPECTRAL REGULARIZATION EXPERIMENTS

The two pathological regimes admit complementary spectral interventions. For each we apply the regularizer consistently to the training Gram, the train–test cross-covariance, and the test self-covariance, and ablate the single tuning knob via held-out NLL.

Shrinkage rescues constant collapse. Define the modified kernel $K'_q(x, x') = (1 - \alpha)\, K_q(x, x') + \alpha\, t\, \delta(x, x')$ where $t = \mathrm{tr}(K)/n$ and δ is the Kronecker delta. This raises $S/\log n$ toward the useful middle while preserving positive definiteness. Applied to the worst constant-collapse configuration in our M1 sweep (HE, $L = 1$, $s = 0.10$; baseline $S/\log n = 0.15$, NLL $= +10.94$), shrinkage at $\alpha = 0.30$ lifts $S/\log n$ to 0.54 and drops NLL from $+10.94$ to $+0.94$ (Table 10), placing the regularized kernel near the useful-hardness sweet spot of the unconstrained sweep (Figure 16, left).

Truncation only mildly helps Haar concentration. Define K_{trunc} by zeroing all but the top-k eigencomponents of K_{tr} and projecting cross-covariances onto the same top-k subspace. Applied

to a Haar-like configuration (HE, $L = 5$, $s = 1.5$; baseline $S/\log n = 0.998$, NLL $= +1.21$), the best truncation at $k = 15$ yields a marginal NLL improvement to $+1.16$ (Table 10). The lack of strong rescue is consistent with our theoretical claim that information was never injected into the posterior in the first place: simple post-hoc spectral truncation cannot substantially recover it in this experiment, though we do not claim no post-hoc transform whatsoever could. The lesson is operational—for a configuration flagged **near-Haar** by Algorithm 1 (high entropy *and* poor held-out variance contraction), fix the kernel *before* fitting rather than rely on post-hoc spectral truncation.

Table 10: Spectral regularization trajectories. Top: shrinkage on constant-collapse (HE, $L = 1$, $s = 0.10$). Bottom: truncation on Haar-like (HE, $L = 5$, $s = 1.5$). Best NLL row in bold.

α	$S/\log n$	NLL	ECE	VC
0.00	0.151	+10.94	0.377	0.016
0.05	0.236	+2.07	0.213	0.075
0.10	0.307	+1.25	0.160	0.132
0.20	0.432	+0.95	0.123	0.241
0.30	**0.542**	**+0.94**	**0.107**	**0.347**
0.50	0.729	+1.02	0.083	0.551
0.70	0.877	+1.12	0.090	0.747
0.90	0.979	+1.20	0.090	0.933

k	$S/\log n$	NLL	ECE	VC
30 (full)	0.998	+1.21	0.097	0.987
25	0.945	+1.19	0.107	0.986
20	0.880	+1.17	0.113	0.985
15	**0.795**	**+1.16**	**0.137**	**0.983**
10	0.676	+1.22	0.210	0.979
5	0.471	+1.68	0.320	0.967
3	0.321	+2.44	0.373	0.953
1	0.000	+6.45	0.433	0.898

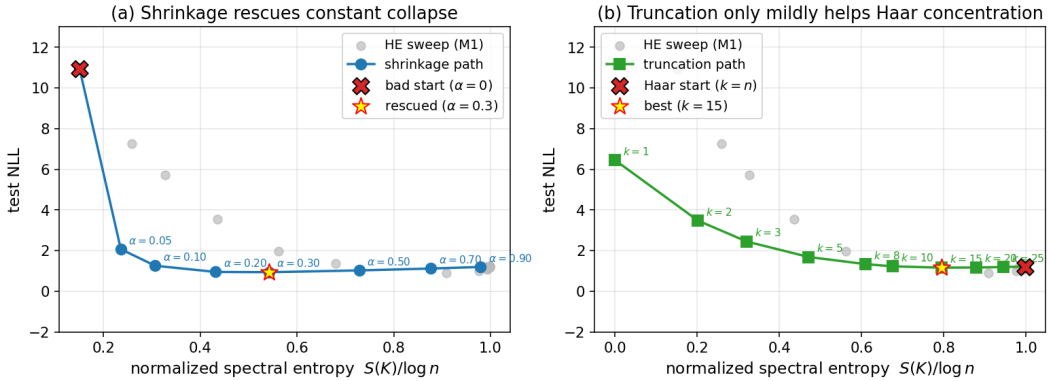


Figure 16: Spectral regularization trajectories on the M1 sweep manifold (gray points). (a) Shrinkage applied to the constant-collapse kernel (red X) sweeps it through the useful-hardness frontier, recovering the sweet-spot NLL at $\alpha \approx 0.30$ (yellow star). (b) Truncation applied to a Haar-like kernel (red X) only mildly improves NLL at $k = 15$, confirming that information loss at Haar concentration is largely irreversible.

AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.