CHARACTERIZING PROJECTED QUANTUM KERNEL LEARNING VIA DATA-TO-READOUT INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantum kernel methods map classical data into high-dimensional quantum state spaces, but projected learners access only a prescribed readout, leaving unclear how circuit structure determines the information available for learning. We introduce the *data-to-readout interface* to characterize this transfer. For linear projected quantum kernels, we show that the reachable readout dimension is exactly determined by the transfer channel restricted to the encoder span, while the intrinsic transfer dimension is bounded by the squared operator Schmidt rank of the fixed quantum process. Controlled interfaces further admit an exact factorization into input-dependent probabilities and fixed readout geometry. This factorization characterizes the population-kernel spectrum and effective dimension, determines when widening the interface provides additional target-relevant information, and yields finite-sample and measurement guarantees. We further show that statistical simplicity does not imply classical tractability: under standard complexity and cryptographic assumptions, even low-width interfaces can remain classically hard to evaluate, and a logarithmic-width construction yields a quantum–classical learning separation. Experiments on controlled circuits with up to 1000 qubits and on disordered transverse-field Ising ground states validate the structural bounds, target-alignment predictions, and finite-measurement trade-offs.

1 INTRODUCTION

Quantum computing exploits high-dimensional state spaces and coherent dynamics to process information in ways that differ from classical computation (Nielsen and Chuang, 2000), with applications in quantum simulation (Feynman, 1982), scientific computing (Harrow et al., 2009), and cryptanalysis (Shor, 1997). These capabilities have also motivated quantum machine learning for classification (Rebentrost et al., 2014), regression (Schuld et al., 2016), and related prediction tasks. A quantum feature map can embed classical inputs into an exponentially large Hilbert space (Schuld and Killoran, 2019), but the learner never accesses that state space directly. Prediction depends only on the information that reaches the prescribed measurements at the readout (Schuld, 2021). The relevant capacity of a quantum model is therefore set not only by the representation created inside the circuit, but also by the route through which encoded information reaches the learner.

Quantum kernel methods (Mengoni and Di Pierro, 2019) provide a particularly transparent setting for studying this route. Full-state kernels use overlaps between states produced by a quantum feature map (Havlíček et al., 2019), yet a large feature space is not automatically useful. As the effective dimension grows, pairwise similarities can concentrate (Thanasilp et al., 2024); the resulting spectrum can demand more training samples (Kübler et al., 2021) and more quantum measurements (Miroszewski et al., 2024). Projected quantum kernels (Huang et al., 2021) reduce this burden by restricting learning to reduced states or other local quantum features (Gan et al., 2023). Existing analyses mainly begin after this projected representation has been specified, studying its spectrum (Canatar et al., 2023), generalization (Huang et al., 2021), or measurement cost (Gentinetta et al., 2024). What remains less explicit is the circuit-level map from encoded data to the operator space visible at a particular readout. This motivates three related questions:

How much input information can reach the readout, which part is useful under finite learning resources, and when can the resulting representation be evaluated classically?

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.

We study these questions through the *data-to-readout interface*, the circuit region through which input dependence reaches the measured subsystem. The first quantity is the *reachable dimension* D, the dimension of the span of readout states generated on the support of the data distribution. We then define the *intrinsic transfer dimension* r_{tr} as the rank of the fixed channel that transfers operators from the encoding register to an initially blank readout. Theorem 1 shows $D \le r_{\mathrm{tr}} \le r^2$, where r is the operator Schmidt rank of the fixed entangling process across the encoder–readout partition. The hierarchy separates three effects that are otherwise conflated: which directions are activated by the data, how many directions the fixed process can transfer, and how the circuit constrains that transfer.

For controlled interfaces, the same structure becomes explicit. Corollary 1 gives $\rho_B(x) = \sum_s p_s(x)\tau_s$ and $k_B(x,x') = p(x)^\top G p(x')$, with $D \le \mathrm{rank}(G) \le 2^{b_{\mathrm{eff}}} \le 2^b$. Here the probabilities $p(x)$ carry all input dependence, while the reusable matrix G records the geometry of fixed readout branches. This factorization lets us move from structural capacity to learning. Corollary 3 identifies the population-kernel spectrum and effective dimension from a finite matrix, gives the best squared-loss approximation error in the reachable span, and characterizes when a wider interface contributes target information that is absent from a narrower one. The statistical results in Appendices E and F then quantify how finite samples and finite measurements determine which of these directions can actually be exploited.

We next separate statistical simplicity from computational accessibility. Theorem 2 shows that a controlled interface with polynomially many coordinates is classically tractable when both its input-dependent probabilities and reusable branch geometry can be approximated classically in polynomial time. Low dimension by itself gives no such guarantee. Assuming $\mathrm{BQP} \not\subseteq \mathrm{BPP}$, constant-dimensional interfaces can still contain classically hard quantities. Under an injective function that is classically one-way but quantumly invertible, we further construct a logarithmic-width interface with $D = \Theta(n)$ whose unknown target is identified quantumly from $O(\log n)$ ordinary labeled examples, while every probabilistic polynomial-time classical learner has only negligible average prediction advantage from polynomially many such examples. Corollary 4 shows that established quantum-learning separations also admit one-qubit or logarithmic-width projected realizations.

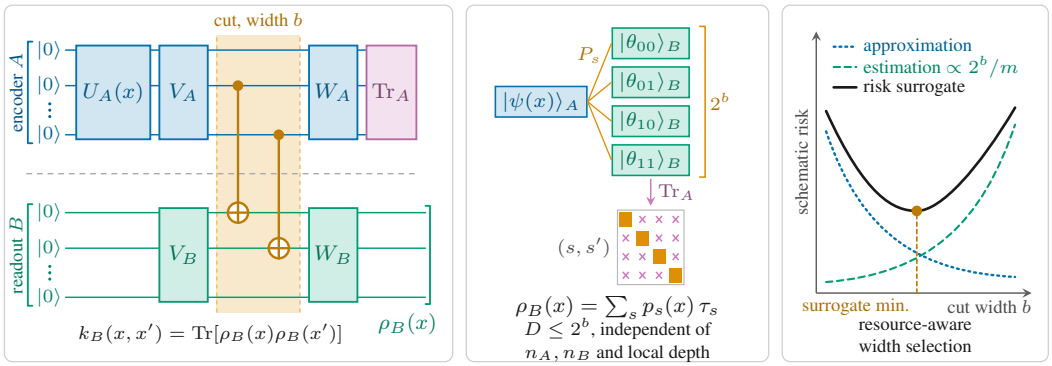
The experiments test the distinctions that remain after the theory fixes the structural and statistical quantities. Controlled circuits with up to 1000 qubits attain the predicted capacity ceilings, and a balanced family follows the D/m finite-sample risk scaling. On disordered transverse-field Ising ground states, widening the interface helps sharply only when the added readout access captures information relevant to the target. When the relevant qubits and local measurement bases are hidden, validation-based selection recovers the exact support and bases in 91.7% of trials and reaches near-oracle prediction error. Finite-shot experiments expose a second gap between structural and usable capacity: at width $b = 8$, only a median of 3 out of 256 structural directions are resolved even with 10^5 SWAP shots per Gram entry. A fixed state-preparation budget also exhibits the predicted sample–measurement trade-off.

In summary, the data-to-readout interface provides a single circuit-level object from which three questions can be asked in order: what representation can reach the readout, what part of it is useful for the target under finite resources, and whether those useful coordinates are classically accessible. The structural, statistical, and computational answers need not coincide, and the experiments isolate each distinction.

2 Preliminaries

We summarize the quantum and kernel concepts used in our analysis. Additional notation, learning formulations, and extended related work are collected in Appendix A.

Basics of quantum computing. A pure qubit state is a unit vector $|\psi\rangle = a|0\rangle + b|1\rangle$ with $a, b \in \mathbb{C}$ and $|a|^2 + |b|^2 = 1$, while an n-qubit state lies in $(\mathbb{C}^2)^{\otimes n}$. General states are density matrices $\rho \succeq 0$ with $\mathrm{Tr}\,\rho = 1$, and a pure state is represented by $\rho = |\psi\rangle\langle\psi|$ (Nielsen and Chuang, 2000). Quantum gates are unitary operators. A controlled unitary from control qubit a to target register β has the form $\mathrm{CU}_{a\to\beta} = |0\rangle\langle 0|_a \otimes I_\beta + |1\rangle\langle 1|_a \otimes U_\beta$, with the CNOT gate obtained by taking $U = X$ on a target qubit (Nielsen and Chuang, 2000). For a bipartite state ρ_{AB}, the reduced state on B is $\rho_B = \mathrm{Tr}_A\,\rho_{AB}$, which preserves the statistics of all measurements acting only on B. For an operator U on $A \otimes B$, an operator Schmidt decomposition writes $U = \sum_{j=1}^{r} \sigma_j A_j \otimes B_j$ with orthogonal



(a) Bottleneck circuit **(b)** Which-path cancellation **(c)** Capacity–sample tradeoff

Figure 1: **The data-to-readout interface limits the representation exposed to the learner. (a)** Data enter encoder A and reach readout B only through the interface. **(b)** For a controlled interface, partial trace removes cross-sector terms and leaves a mixture over fixed readout branches. **(c)** Increasing interface width can improve approximation but also increases the worst-case finite-sample estimation cost, schematically scaling as $2^b/m$, which motivates resource-aware width selection. The curves are schematic.

local operators A_j and B_j. The minimum number of terms, $r = \mathrm{SR}_{A:B}(U)$, is the operator Schmidt rank (Nielsen et al., 2003). Local gates have rank 1, controlled unitaries have rank at most 2, and two-qubit gates crossing the partition have rank at most 4.

Quantum kernel methods. A positive semidefinite kernel $k(x, x')$ defines pairwise similarities between inputs, with Gram matrix $K_{ij} = k(x_i, x_j)$ on a sample $\{x_i\}_{i=1}^m$ (Schölkopf and Smola, 2001). Quantum feature maps encode x into a quantum state $\rho(x)$ (Schuld and Killoran, 2019), and state overlaps can be used as kernel entries (Havlíček et al., 2019). We study the linear Hilbert–Schmidt kernel $k(x, x') = \mathrm{Tr}[\rho(x)\rho(x')]$. For a projected quantum kernel, the learner observes only subsystem B (Huang et al., 2021), giving $\rho_B(x) = \mathrm{Tr}_A\,\rho(x)$ and $k_B(x, x') = \mathrm{Tr}[\rho_B(x)\rho_B(x')]$. This linear form is central because its feature span and spectrum can be related directly to circuit structure. More general projected and trace-induced constructions may apply nonlinear kernels to the same reduced representation (Gan et al., 2023), but such post-processing cannot recover input dependence that never reaches $\rho_B(x)$.

Related work. Projected quantum kernels restrict learning to reduced quantum representations (Huang et al., 2021), while subsequent studies have analyzed quantum-kernel inductive bias and spectrum (Kübler et al., 2021), target alignment (Canatar et al., 2021), bandwidth (Shaydulin and Wild, 2022), and feature dimension (Kempkes et al., 2026). Trace-induced kernels generalize kernels constructed from reduced quantum states (Gan et al., 2023), while active quantum subspaces study statistically useful dimensions defined through measured observables (Bang et al., 2026). These works largely characterize a representation after the readout or measured features have been specified. Circuit-level analyses instead constrain the model before this stage, for example through trainable gate counts (Caro et al., 2022) or concentration induced by deeper data encoding (Li et al., 2022). Other work studies when quantum learning models admit efficient classical surrogates or classical simulation (Schreiber et al., 2023; Cerezo et al., 2025; Gil-Fuster et al., 2025), quantum–classical learning separations under different access assumptions (Servedio and Gortler, 2004; Molteni et al., 2026; Yamasaki et al., 2026), and limitations imposed by finite measurements (Wang et al., 2021; Thanasilp et al., 2024). Our analysis connects these stages by asking how circuit structure determines the input-dependent operator span that reaches a prescribed readout.

3 Data-to-Readout Interface

A projected quantum kernel is built from the reduced representation $x \mapsto \rho_B(x)$ on a measured subsystem B. We study the circuit mechanism that creates this representation. The central object is

the *data-to-readout interface*: the part of the fixed circuit through which input dependence passes from the encoding register to the readout. Figure 1 summarizes the three levels of the analysis. We first bound the readout directions that the interface can transmit, then determine which of them are populated and target-relevant under a data distribution, and finally ask whether the resulting coordinates are classically accessible. Complete proofs are grouped in Appendices B to F.

3.1 Interface Representation and Structural Capacity

Let P denote the input distribution. Split the circuit into an encoding register A and a readout register B initialized in $|0\rangle_B$. Data enter only through $U_A(x)$, after which a fixed unitary U_{ent} produces $|\Phi(x)\rangle = U_{\mathrm{ent}}(U_A(x)|0\rangle_A \otimes |0\rangle_B)$ and the learner observes $\rho_B(x) = \mathrm{Tr}_A|\Phi(x)\rangle\langle\Phi(x)|$. For the linear projected kernel $k_B(x, x') = \mathrm{Tr}[\rho_B(x)\rho_B(x')]$, the first quantity of interest is

$$D = \dim \mathrm{span}\{\rho_B(x) : x \in \mathrm{supp}(P)\}. \tag{1}$$

The ambient operator space on an n_B-qubit readout has dimension 4^{n_B}, whereas D counts only the independent readout directions that are actually reachable on the data support.

The circuit imposes a second restriction before the data distribution is taken into account. Let $\rho_A(x) = U_A(x)|0\rangle\langle 0|_A U_A(x)^\dagger$ and $\mathcal{V}_A = \mathrm{span}\{\rho_A(x) : x \in \mathrm{supp}(P)\}$. The fixed process and blank readout define the linear transfer map $\mathcal{N}_{A\to B}(X) = \mathrm{Tr}_A[U_{\mathrm{ent}}(X \otimes |0\rangle\langle 0|_B)U_{\mathrm{ent}}^\dagger]$. We call $r_{\mathrm{tr}} = \mathrm{rank}(\mathcal{N}_{A\to B})$ on the real Hermitian-operator spaces the *intrinsic transfer dimension*, and write $r = \mathrm{SR}_{A:B}(U_{\mathrm{ent}})$ for the operator Schmidt rank of the fixed unitary. The following theorem relates the data-dependent representation to these two circuit quantities.

Theorem 1 (Interface capacity)**.** *The reachable readout dimension satisfies*

$$D = \mathrm{rank}(\mathcal{N}_{A\to B}|_{\mathcal{V}_A}) \le r_{\mathrm{tr}} \le r^2. \tag{2}$$

If a realization of U_{ent} contains b gates crossing the A:B cut and each has operator Schmidt rank at most κ, then $r \le \kappa^b$.

The three quantities in Eq. (2) answer different questions. D depends on the encoder and the support of P; r_{tr} is a property of the fixed transfer channel induced by U_{ent} and the blank readout; and r depends only on the bipartite unitary. If $\mathcal{V}_A$ spans the full Hermitian operator space on A, then $D = r_{\mathrm{tr}}$. At the circuit level, any gate decomposition provides an upper bound on r from the number and type of crossings. The optimized crossing-count refinement and its relation to tensor-network bond dimension are given in Appendix B.

A controlled interface gives a sharper and more explicit representation. In particular, suppose

$$U_{\mathrm{ent}} = (W_A \otimes W_B)\left(\prod_{i=1}^{b} \mathrm{CU}_{a_i\to\beta_i}\right)(V_A \otimes V_B). \tag{3}$$

Let S be the set of distinct control qubits, $b_{\mathrm{eff}} = |S|$, and P_s the projector onto control sector $s \in \{0,1\}^S$. After absorbing V_A into the encoder, write $|\psi(x)\rangle = V_A U_A(x)|0\rangle_A$ and $p_s(x) = \langle\psi(x)|P_s|\psi(x)\rangle$; collect these probabilities in $p(x)$. The partial trace removes coherences between different control sectors, which yields the following factorization.

Corollary 1 (Controlled-interface factorization)**.** *Under Eq. (3), there are fixed readout states $\{\tau_s\}$ such that*

$$\rho_B(x) = \sum_s p_s(x)\tau_s, \qquad k_B(x,x') = p(x)^\top G p(x'), \qquad D \le \mathrm{rank}(G) \le 2^{b_{\mathrm{eff}}} \le 2^b, \tag{4}$$

where $G_{ss'} = \mathrm{Tr}(\tau_s\tau_{s'})$.

This decomposition separates input-dependent coordinates from readout geometry. The vector $p(x)$ is acquired for each input, whereas G depends only on the fixed branch states and can be reused across the data set. The rank of G is therefore a circuit-dependent ceiling, not the dimension necessarily occupied by a particular data distribution: the latter is still D and can be smaller because the vectors $p(x)$ may explore only a subspace of the available sectors. Equation equation 4 also separates three notions of width that need not agree. The raw crossing count b is a property of a

chosen circuit realization, b_{eff} counts distinct control bits, and $\text{rank}(G)$ records linear dependence among the fixed readout branches. Increasing the readout register without changing these branches creates no new feature direction, while different branch geometries at the same b can yield different ceilings. Repeated crossings controlled by the same qubit share a sector label, so the relevant control-side width is b_{eff} rather than the raw number of crossings. The same factorization remains valid with fixed sector-preserving operations interleaved between controlled gates; precise conditions and target-side refinements are given in Appendices B and C.

The physical $A{:}B$ cut need not be the narrowest point of the circuit. Suppose instead that a q-qubit subsystem C intercepts every input-dependent path to a retained readout R, and that the circuit after this cut is an input-independent channel $\mathcal{E} : C \to R$. Let D_R denote the span dimension of the retained readout states. Calling C a *causal separator* gives a bound independent of the final readout size.

Corollary 2 (Causal-separator ceiling)**.** *For a q-qubit causal separator, $D_R \le \text{rank}(\mathcal{E}) \le 4^q$. If the separator is classical with L possible branches, then $D_R \le L$.*

This form is useful when local simplification or light-cone reduction exposes a bottleneck away from the final encoder–readout boundary. The remainder of the main analysis focuses on the controlled form, for which the transferred coordinates are explicit.

3.2 Effective Capacity and Learning Resources

Structural dimension alone does not determine learning performance. The data distribution determines how strongly the reachable directions are populated, and the target determines whether those directions are useful. For squared-loss regression, let $f^\circ(x) = \mathbb{E}[y \mid x]$. Assume $x \mapsto \rho_B(x)$ is continuous on $\text{supp}(P)$ and $\mathbb{E}[f^\circ(x)^2] < \infty$. For the controlled factorization, define $C_p = \mathbb{E}[p(x)p(x)^\top]$, $z(x) = G^{1/2}p(x)$, $M = G^{1/2}C_pG^{1/2}$, and $c = \mathbb{E}[z(x)f^\circ(x)]$. The finite matrix $M = \mathbb{E}[z(x)z(x)^\top]$ contains the population spectrum seen by the linear projected kernel.

Corollary 3 (Population spectrum and target utility)**.** *Under the assumptions above, the nonzero population-kernel spectrum is exactly the nonzero spectrum of M. Consequently,*

$$D = \text{rank}(M), \qquad d_{\text{eff}}(\gamma) = \text{Tr}[M(M+\gamma I)^{-1}], \qquad \mathcal{A} = \mathbb{E}[f^\circ(x)^2] - c^\top M^\dagger c. \tag{5}$$

Here $\mathcal{A}$ is the best population squared-loss approximation error in the feature span. For a nested enlargement that appends features $u(x)$ to $z(x)$, the optimal ridge objective improves exactly when the added coordinates correlate with the residual of the narrower ridge solution. Writing $C_{zu} = \mathbb{E}[z(x)u(x)^\top]$ and $c_u = \mathbb{E}[u(x)f^\circ(x)]$, this condition is $c_u - C_{zu}^\top(M+\gamma I)^{-1}c \neq 0$.

The rank of M counts directions with nonzero population variance, while $d_{\text{eff}}(\gamma)$ discounts directions that are weak at the regularization scale γ. Target utility introduces a separate quantity: $\mathcal{A}$ depends on c, so two interfaces with the same D and similar spectra can have different prediction error. Thus the structural question and the learning question are deliberately separated: D specifies what can vary on the data support, $d_{\text{eff}}(\gamma)$ specifies what is statistically resolved at a chosen scale, and $\mathcal{A}$ asks whether the resolved span contains the target. For a nested pair, the residual criterion states exactly when the added coordinates contribute information not already captured by the narrower ridge solution. The coordinate-free version for general nesting maps is Proposition 2 in Appendix C; held-out selection over a finite interface family is analyzed in Proposition 4 of Section E.1.

For a sample $\{x_i\}_{i=1}^m$, let $Q_{is} = p_s(x_i)$. The empirical Gram matrix factorizes as

$$K = QGQ^\top. \tag{6}$$

Each row of Q contains the interface statistics of one input, while G is shared across the data set. Once G is available, a new input requires only one additional row of Q. Equivalently, the explicit features $z(x) = G^{1/2}p(x)$ allow training and prediction directly in the interface dimension instead of through an $m \times m$ Gram matrix.

Finite labeled data impose a statistical resolution scale on this population representation. Under the regularized covariance-resolution and noise conditions stated in Section E.3, Proposition 6 gives a KRR estimation term of order $\{d_{\text{eff}}(\gamma) + \log(1/\delta)\}/m$, in addition to approximation and ridge-bias

terms. Since $d_{\text{eff}}(\gamma) \le D \le 2^b$ for the controlled interface, the rank envelope gives the coarse scaling $2^b/m$, whereas the spectral form can be much smaller when weakly populated directions are unresolved at scale γ.

Quantum measurements impose a distinct resource. When G is known, measuring the control basis samples $p(x)$ directly, and substituting empirical probabilities into Eq. (6) preserves positive semidefiniteness. Uniform entrywise Gram accuracy ε over m inputs can be achieved with $O(m\varepsilon^{-2}\log(m/\delta))$ interface-state preparations, while a two-branch subclass requires $\Omega(m\varepsilon^{-2})$ preparations; the matching statements are Propositions 12 and 13 in Section F.4. Thus additional labeled inputs and repeated measurements reduce different errors. The joint KRR bound in Proposition 7 of Section E.4 makes this separation explicit and yields an interior sample–shot balance under a fixed preparation budget.

3.3 Computational Accessibility and Quantum Learning

A low-dimensional representation can still be expensive to compute. In the controlled factorization, a classical implementation needs numerical access to both the input-dependent vector $p(x)$ and the reusable matrix G. The next theorem identifies an efficient regime and then shows why dimension alone cannot certify it.

Theorem 2 (Computational and learning frontier)**.** *Suppose the controlled interface has polynomially many coordinates. If $p(x)$ is classically computable to inverse-polynomial ℓ_1 accuracy and G to inverse-polynomial entrywise accuracy in polynomial time, then its Gram matrix, and hence any polynomial-time downstream kernel procedure, is classically computable to inverse-polynomial precision.*

Assuming BQP $\not\subseteq$ BPP*, constant-dimensional interfaces exist for which either the input-dependent probabilities $p(x)$ or a reusable entry of G is BQP-hard to approximate classically. Moreover, for a family indexed by input size n, if there exists an injective polynomial-time computable function that is classically one-way but quantumly invertible, there is a controlled realization with $b = O(\log n)$, $G = I$, and $D = \Theta(n)$ whose unknown target is identified quantumly from $O(\log(n/\delta))$ ordinary labeled examples, whereas every probabilistic polynomial-time classical learner has only negligible average prediction advantage from polynomially many such examples.*

The efficient statement follows directly from Eq. (6): approximating every row of Q and the reusable entries of G is sufficient to approximate the full polynomial-size Gram matrix. The two constant-dimensional constructions place hardness on opposite sides of the factorization. One keeps $G = I$ and encodes the hard quantum computation in $p(x)$; the other makes the input probabilities trivial and places the hard quantity in a single branch-overlap entry of G. The end-to-end learning construction is stronger than either evaluation witness. Its public dictionary has $\Theta(n)$ independent interface directions but only logarithmic crossing width; a quantum learner identifies the hidden target from logarithmically many labels, while a classical learner with polynomially many labels would otherwise predict a Goldreich–Levin hard-core bit with non-negligible advantage. The reductions and quantifiers are given in Sections D.1 and D.2.

The same interface analysis can also be applied to existing quantum-learning separations to determine the projected capacity they require.

Corollary 4 (Narrow interfaces can retain end-to-end learning separations)**.** *Two established quantum learning separations admit statistically small projected realizations. Under the discrete-logarithm hardness assumption of Liu et al. (2021a), their concept class has a one-qubit causal-separator realization with $D = 3$. Under the assumption and observable family of Molteni et al. (2026, Theorem 15), the corresponding identification separation admits a controlled projected realization with $b_{\text{eff}} = O(\log n)$ and $D = \text{poly}(n)$.*

The cited works supply the corresponding quantum learners and classical lower bounds; our role is to expose how much projected readout capacity their target families need. This distinction matters because hardness of a single kernel entry and hardness of learning are not interchangeable. The constant-dimensional witnesses rule out an inference from low capacity to easy classical evaluation, whereas the logarithmic-width construction and the inherited separations concern the end-to-end learning task itself. Together with Theorem 2, these realizations show that a statistically narrow readout can coexist with computationally hard feature generation.

4 EXPERIMENTS

We test the main consequences of the interface analysis: whether the structural ceilings are attainable, whether the finite-sample scaling appears at practical budgets, whether widening helps only when it exposes target-relevant information, and how finite measurements reduce the usable capacity. The experiments deliberately separate these questions. Synthetic controlled circuits isolate the structural and finite-sample predictions, disordered TFIM ground states test target alignment and interface selection on a quantum many-body task, and MNIST-based controlled circuits probe finite-shot spectral resolution and acquisition cost. Shared circuits, data generation, learning protocols, and numerical rank diagnostics are specified in Appendix G; complete experimental grids and uncertainty estimates are reported in Appendix H.

4.1 STRUCTURAL CAPACITY AND RISK SCALING

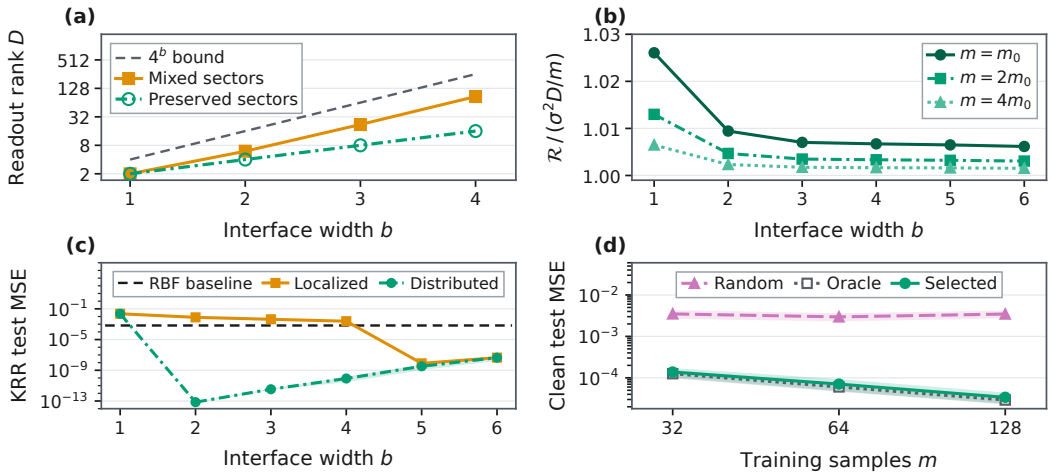


Figure 2: **Structural bounds, risk scaling, and target alignment.** **(a)** Numerical readout ranks under sector-preserving and sector-mixing gates between identical CNOT crossings. **(b)** Exact expected KRR excess risk normalized by $\sigma^2 D/m$, where σ^2 is the label-noise variance; Section H.2.3 defines m_0. **(c)** TFIM regression for $\langle Z_0 Z_4 \rangle$ under two wirings, with an RBF baseline on Hamiltonian coefficients. **(d)** Held-out selection among 1120 width-4 interfaces with target support and bases hidden. Bands in (c,d) are ± 1.96 standard errors over ten splits and 24 tasks, respectively.

For the tested interfaces with distinct controls, the common-sector factorization in corollary 1 gives $D \leq 2^b$. If local gates mix those control sectors between crossings, the controlled mixture no longer applies; Proposition 8 in Section F.1 instead gives the sharper interleaved bound $D \leq (4^b + 2)/3$. Holding the CNOT crossings fixed and changing only the intervening local gates, the sector-preserving circuits attain ranks $2, 4, 8, 16$, whereas the sector-mixing circuits attain $2, 6, 22, 86$ for $b = 1, \ldots, 4$ (Figure 2(a)). The two sequences exactly attain their respective ceilings, isolating the effect of sector mixing at a fixed crossing budget. Full rank diagnostics and seed checks are in Sections H.1 and H.1.2.

The common-sector ceiling also saturates in all 30 large-register configurations with $N \in \{20, 100, 200, 500, 1000\}$ and $b = 0, \ldots, 5$, evaluated through $K = QGQ^\top$. At $N = 1000$ and $b = 5$, both Q and G have rank 32, showing that neither side of the factorization is responsible for an accidental rank loss (Section H.1.1).

We next isolate the finite-sample term. The rank-based KRR result Proposition 5 in Section E.2 predicts a D/m estimation contribution when the populated covariance is resolved and the remaining quantities are held fixed. We construct a controlled family with $D = 2^b$ and population covariance I_D/D, for which these conditions can be checked exactly. Across 18 configurations, the exact expected KRR excess risk is 1.002–1.026 times $\sigma^2 D/m$ (Figure 2(b); Section H.2.3). This experiment tests the scaling in a balanced family rather than the universal constant in the general upper bound.

4.2 Target Alignment and Interface Selection

The residual criterion in corollary 3, proved in Section C.1, predicts that widening is useful only when the newly exposed coordinates contain target information not already captured by the narrower ridge solution. We test this on 240 exactly computed ground states of disordered eight-qubit transverse-field Ising Hamiltonians (Pfeuty, 1970). For the target $\langle Z_0 Z_4 \rangle$, widening a localized interface from $b = 1$ to 4 reduces test MSE by $9.3\times$, but the error remains $3.6\times$ above the RBF baseline at $b = 4$. The target becomes exactly representable at $b = 5$, where the MSE reaches 7.98×10^{-9}. A distributed wiring exposes both target qubits already at $b = 2$ and reaches 7.06×10^{-14} (Figure 2(c)). The intermediate widths still benefit from indirect correlations, but the largest improvement occurs when the relevant coordinates themselves enter the readout (Figure 4 in Section H.3).

We then hide both the support and local measurement bases. Each of 24 tasks combines two disjoint two-body Pauli correlations, while the selector sees only noisy labels and 56 validation states. It jointly chooses ridge regularization and one of $\binom{8}{4}2^4 = 1120$ width-4 interfaces. For $m = 32, 64, 128$, the selected models approach the support-aware oracle and reduce clean test MSE by 25–$102\times$ relative to random interface selection. Exact support-and-basis recovery is $22/24$ (91.7%) at every sample size (Figure 2(d); Section H.3.1). The result concerns selection within this prescribed candidate family; the held-out selection guarantee used to motivate the procedure is Proposition 4 in Section E.1.

4.3 Finite Measurements and Acquisition Cost

The factorization in Eq. (6) separates per-input acquisition of Q from reusable calibration of G, while the learning bounds separate measurement error from finite-sample error. We therefore evaluate three corresponding effects: how samples and shots should be allocated under a fixed preparation budget, how many structural directions remain spectrally visible after finite-shot kernel estimation, and how much reusable interface structure reduces the cost of acquiring or calibrating the kernel.

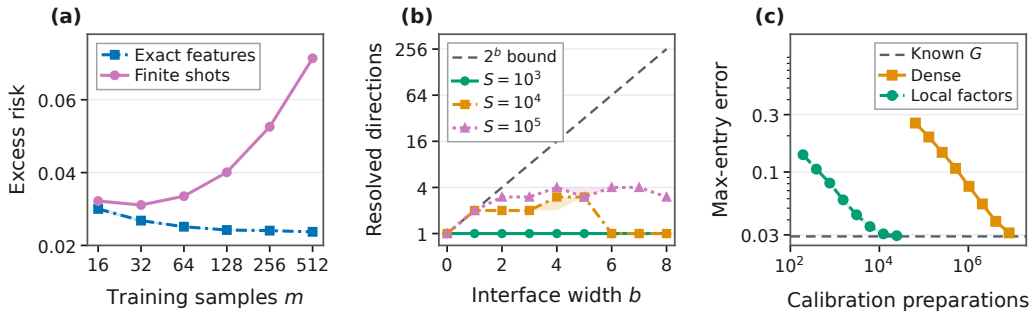


Figure 3: **Finite-shot resolution and measurement costs.** **(a)** Mean population excess risk over the full sample-budget sweep at $mS = 4096$, with ± 1.96 standard errors over 2000 training realizations. **(b)** Resolved directions under independent SWAP sampling: medians and 5–95% ranges over 50 acquisitions, with S shots per Gram entry. **(c)** Mean maximum-entry kernel error with ± 1.96 standard errors over 200 repeats at $b = 6$; costs count calibration preparations only.

Sample–shot allocation. The joint bound Proposition 7 in Section E.4 contains separate finite-sample and finite-measurement terms, so a fixed preparation budget need not be optimized by maximizing the number of labeled inputs. At $mS = 4096$, the exact-feature risk decreases from 0.0300 at $m = 16$ to 0.0237 at $m = 512$. With S-shot training and query features, the mean risk instead reaches its minimum 0.0311 at $(m, S) = (32, 128)$ and rises to 0.0714 at $(512, 8)$ (Figure 3(a); Table 13 in Section H.4.9). Query measurements are excluded from this training budget.

Spectral resolution. The finite-shot criterion Proposition 11 in Section F.3 distinguishes structural rank from directions resolvable above the matrix-Bernstein perturbation scale (Tropp, 2015). On the 300-input Gram block from the fixed MNIST-derived representation described in Appendix G, at $b = 8$ and $S = 10^5$ SWAP shots per entry, only a median of 3 directions out of structural rank 256 are resolved; $49/50$ acquisitions resolve exactly 3 (Figure 3(b)). This experiment uses MNIST

Table 1: **Kernel acquisition at matched preparation budgets.** Mean relative Frobenius errors over 100 repeats for n inputs and budget $n(n+1)T$, with $T = 4$ SWAP shots/pair or $S = (n+1)T$ interface samples/input. The interface estimator knows G; the best baseline is selected from SWAP, random-Pauli, and local randomized-measurement (RM) estimators, each PSD-projected and rank-truncated. Full grid and standard errors: Section H.4.4.

n	b	Best baseline	Relative Frobenius error ↓		Error ratio ↑
		protocol	Baseline	Interface	Baseline / interface
32	2	Local RM	0.2818	**0.0490**	5.8×
	4	Local RM	1.5256	**0.0815**	18.7×
	6	SWAP	3.5354	**0.1134**	31.2×
	8	SWAP	5.0153	**0.1359**	36.9×
128	2	Local RM	0.1168	**0.0263**	4.4×
	4	Local RM	0.7079	**0.0409**	17.3×
	6	SWAP	4.2573	**0.0545**	78.2×

to supply a fixed, nontrivial feature distribution for spectral-resolution analysis rather than as an inductive image-classification benchmark. The spectral-edge diagnostic in Section H.4.1 estimates 1.6×10^{24} shots per entry for the weakest structural direction.

Acquisition and calibration. Table 1 compares interface sampling through $K = QGQ^\top$ with SWAP and two reusable randomized-measurement protocols at equal state-preparation budgets and under the same PSD and rank constraints. Across 19 configurations, known G reduces mean relative Frobenius error by 4.44–78.19× relative to the best baseline in each configuration; the complete comparison is reported in Section H.4.4.

When G is unknown, its reusable structure can still reduce calibration cost. At $b = 6$, dense calibration uses 8,257,536 branch-state preparations for maximum-entry error 0.03115, whereas local-factor calibration uses 24,576 preparations for error 0.02951, a 336× reduction at comparable accuracy (Figure 3(c); Section H.4.6). Both additionally use 32×2048 interface preparations per kernel build. This comparison concerns state-preparation counts for the specified estimators; equal counts need not imply equal device time or a general quantum advantage (Bowles et al., 2024; Schnabel and Roth, 2025).

5 Conclusion

The data-to-readout interface identifies the operator space that a projected quantum learner can actually access. The transfer map separates this reachable space from the ambient readout dimension, while the controlled-interface factorization makes the separation explicit through input-dependent probabilities and reusable readout geometry. Once the data distribution and target are introduced, the same factorization determines the populated spectrum, effective dimension, and approximation error. Finite samples and measurements then impose further resolution scales. The experiments recover these distinctions directly: both structural ceilings are attainable, the balanced family follows the predicted D/m risk scaling, target-relevant wiring matters more than width alone, and finite shots can leave most structurally available directions unresolved.

The computational results add a separate boundary. Small D or small d_{eff} can make a representation statistically simple without making its coordinates classically easy to obtain. Classical tractability follows when both sides of the interface factorization are efficiently approximable, whereas the constant-dimensional witnesses and the logarithmic-width hard-core construction show that this condition is not implied by low statistical capacity. In this sense, interface width is an architectural constraint rather than a stand-alone criterion for quantum advantage.

The present analysis is sharpest for linear projected kernels with a prescribed readout and a fixed post-encoding process. Nonlinear kernels can change the downstream geometry but cannot recover input dependence that never reaches the reduced representation. Trainable interfaces, adaptive measurements, and data re-uploading fall outside this fixed-transfer setting; extending the analysis to them and weakening the computational assumptions are natural next steps.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with language editing, alternative formulations, proof exposition and consistency checks, and LaTeX presentation. The research questions, mathematical claims, experimental data, analyses, interpretations, and conclusions were developed by the authors. All AI-assisted text and proof edits were independently checked against the underlying derivations, and any suggested factual statements or citations were verified by the authors. The authors take responsibility for the final manuscript.

REFERENCES

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

Richard P. Feynman. Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6–7):467–488, 1982. doi: 10.1007/BF02650179.

Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical Review Letters*, 103(15):150502, 2009. doi: 10.1103/PhysRevLett.103.150502.

Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5):1484–1509, 1997. doi: 10.1137/S0097539795293172.

Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Physical Review Letters*, 113(13):130503, 2014. doi: 10.1103/PhysRevLett.113.130503.

Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. Prediction by linear regression on a quantum computer. *Physical Review A*, 94(2):022342, 2016. doi: 10.1103/PhysRevA.94.022342.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.

Maria Schuld. Supervised quantum machine learning models are kernel methods. *arXiv preprint arXiv:2101.11020*, 2021.

Riccardo Mengoni and Alessandra Di Pierro. Kernel methods in quantum machine learning. *Quantum Machine Intelligence*, 1(3):65–71, 2019. doi: 10.1007/s42484-019-00007-4.

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.

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

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

Artur Miroszewski, Marco Fellous Asiani, Jakub Mielczarek, Bertrand Le Saux, and Jakub Nalepa. In search of quantum advantage: Estimating the number of shots in quantum kernel methods. *arXiv preprint arXiv:2407.15776*, 2024.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1):2631, 2021.

Beng Yee Gan, Daniel Leykam, and Supanut Thanasilp. A unified framework for trace-induced quantum kernels. *arXiv preprint arXiv:2311.13552*, 2023.

Abdulkadir Canatar, Evan Peters, Cengiz Pehlevan, Stefan M. Wild, and Ruslan Shaydulin. Bandwidth enables generalization in quantum kernel models. *Transactions on Machine Learning Research*, 2023.

Gian Gentinetta, Arne Thomsen, David Sutter, and Stefan Woerner. The complexity of quantum support vector machines. *Quantum*, 8:1225, 2024. doi: 10.22331/q-2024-01-11-1225.

Michael A. Nielsen, Christopher M. Dawson, Jennifer L. Dodd, Alexei Gilchrist, Duncan Mortimer, Tobias J. Osborne, Michael J. Bremner, Aram W. Harrow, and Andrew Hines. Quantum dynamics as a physical resource. *Physical Review A*, 67(5):052301, 2003. doi: 10.1103/PhysRevA.67.052301.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, 2021. doi: 10.1038/s41467-021-23103-1.

Ruslan Shaydulin and Stefan M. Wild. Importance of kernel bandwidth in quantum machine learning. *Physical Review A*, 106(4):042407, 2022.

Marie Kempkes, Aroosa Ijaz, Elies Gil-Fuster, Carlos Bravo-Prieto, Jakob Spiegelberg, Evert van Nieuwenburg, and Vedran Dunjko. Double descent in quantum kernel methods. *PRX Quantum*, 7(1):010312, 2026. doi: 10.1103/cn64-gs6b.

Jeongho Bang, Wooyeong Song, Kyoungho Cho, Taewan Kim, and Yongsoo Hwang. Learning with active quantum subspaces: Scalable hybrid advantage without full quantum data-encoding. *arXiv preprint arXiv:2606.00932*, 2026.

Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13(1):4919, 2022. doi: 10.1038/s41467-022-32550-3.

Guangxi Li, Ruilin Ye, Xuanqiang Zhao, and Xin Wang. Concentration of data encoding in parameterized quantum circuits. In *Advances in Neural Information Processing Systems*, volume 35, pages 19456–19469, 2022.

Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *Physical Review Letters*, 131(10):100803, 2023. doi: 10.1103/PhysRevLett.131.100803.

M. Cerezo, Martin Larocca, Diego García-Martín, N. L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, Eric R. Anschuetz, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? *Nature Communications*, 16(1):7907, 2025. doi: 10.1038/s41467-025-63099-6.

Elies Gil-Fuster, Casper Gyurik, Adrián Pérez-Salinas, and Vedran Dunjko. On the relation between trainability and dequantization of variational quantum learning models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TdqaZbQvdi.

Rocco A. Servedio and Steven J. Gortler. Equivalences and separations between quantum and classical learnability. *SIAM Journal on Computing*, 33(5):1067–1092, 2004. doi: 10.1137/S0097539704412910.

Riccardo Molteni, Simon C. Marshall, and Vedran Dunjko. Quantum machine learning advantages beyond hardness of evaluation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=on2lie43Kl.

Hayata Yamasaki, Natsuto Isogai, and Mio Murao. Advantage of quantum machine learning from general computational advantages. *npj Quantum Information*, 12:125, 2026. doi: 10.1038/s41534-026-01279-y.

Xinbiao Wang, Yuxuan Du, Yong Luo, and Dacheng Tao. Towards understanding the power of quantum kernels in the NISQ era. *Quantum*, 5:531, 2021. doi: 10.22331/q-2021-08-30-531.

Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, 2021a.

Pierre Pfeuty. The one-dimensional Ising model with a transverse field. *Annals of Physics*, 57(1): 79–90, 1970. doi: 10.1016/0003-4916(70)90270-8.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld. Better than classical? the subtle art of benchmarking quantum machine learning models. *arXiv preprint arXiv:2403.07059*, 2024.

Jan Schnabel and Marco Roth. Quantum kernel methods under scrutiny: A benchmarking study. *Quantum Machine Intelligence*, 7(1):58, 2025. doi: 10.1007/s42484-025-00273-5.

Ethan Bernstein and Umesh Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473, 1997. doi: 10.1137/S0097539796300921.

Seth Lloyd. Universal quantum simulators. *Science*, 273(5278):1073–1078, 1996. doi: 10.1126/science.273.5278.1073.

Iulia M. Georgescu, Sahel Ashhab, and Franco Nori. Quantum simulation. *Reviews of Modern Physics*, 86(1):153–185, 2014. doi: 10.1103/RevModPhys.86.153.

Ashley Montanaro. Quantum algorithms: an overview. *npj Quantum Information*, 2(1):15023, 2016. doi: 10.1038/npjqi.2015.23.

Jin-Peng Liu, Herman Øie Kolden, Hari K. Krovi, Nuno F. Loureiro, Konstantina Trivisa, and Andrew M. Childs. Efficient quantum algorithm for dissipative nonlinear differential equations. *Proceedings of the National Academy of Sciences*, 118(35):e2026805118, 2021b. doi: 10.1073/pnas.2026805118.

Craig Gidney and Martin Ekerå. How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits. *Quantum*, 5:433, 2021. doi: 10.22331/q-2021-04-15-433.

Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020.

Takahiro Goto, Quoc Hoan Tran, and Kohei Nakajima. Universal approximation property of quantum machine learning models in quantum-enhanced feature spaces. *Physical Review Letters*, 127 (9):090506, 2021. doi: 10.1103/PhysRevLett.127.090506.

Carsten Blank, Daniel K. Park, June-Koo Kevin Rhee, and Francesco Petruccione. Quantum classifier with tailored quantum kernel. *npj Quantum Information*, 6(1):41, 2020. doi: 10.1038/s41534-020-0272-6.

Sau Lan Wu, Shaojun Sun, Wen Guan, Chen Zhou, Jay Chan, Chi Lung Cheng, Tuan Pham, Yan Qian, Alex Zeng Wang, Rui Zhang, Miron Livny, Jennifer Glick, Panagiotis Kl. Barkoutsos, Stefan Woerner, Ivano Tavernelli, Federico Carminati, Alberto Di Meglio, Andy C. Y. Li, Joseph Lykken, Panagiotis Spentzouris, Samuel Yen-Chi Chen, Shinjae Yoo, and Tzu-Chieh Wei. Application of quantum machine learning using the quantum kernel algorithm on high energy physics analysis at the LHC. *Physical Review Research*, 3(3):033221, 2021. doi: 10.1103/PhysRevResearch.3.033221.

Annie E. Paine, Vincent E. Elfving, and Oleksandr Kyriienko. Quantum kernel methods for solving regression problems and differential equations. *Physical Review A*, 107(3):032428, 2023. doi: 10.1103/PhysRevA.107.032428.

Teppei Suzuki, Takashi Hasebe, and Tsubasa Miyazaki. Quantum support vector machines for classification and regression on a trapped-ion quantum computer. *Quantum Machine Intelligence*, 6(1):31, 2024a. doi: 10.1007/s42484-024-00165-0.

Thomas Hubregtsen, David Wierichs, Elies Gil-Fuster, Peter-Jan H. S. Derks, Paul K. Faehrmann, and Johannes Jakob Meyer. Training quantum embedding kernels on near-term quantum computers. *Physical Review A*, 106(4):042431, 2022. doi: 10.1103/PhysRevA.106.042431.

Jennifer R. Glick, Tanvi P. Gujarati, Antonio D. Córcoles, Youngseok Kim, Abhinav Kandala, Jay M. Gambetta, and Kristan Temme. Covariant quantum kernels for data with group structure. *Nature Physics*, 20(3):479–483, 2024. doi: 10.1038/s41567-023-02340-9.

Yudai Suzuki, Hideaki Kawaguchi, and Naoki Yamamoto. Quantum Fisher kernel for mitigating the vanishing similarity issue. *Quantum Science and Technology*, 9(3):035050, 2024b. doi: 10.1088/2058-9565/ad4b97.

Cong Lei, Yuxuan Du, Peng Mi, Jun Yu, and Tongliang Liu. Neural auto-designer for enhanced quantum kernels. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8htNAnMSyP.

Zhan Yu, Hongshun Yao, Mujin Li, and Xin Wang. Power and limitations of single-qubit native quantum neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 27810–27823, 2022.

Tongyang Li, Shouvanik Chakrabarti, and Xiaodi Wu. Sublinear quantum algorithms for training linear and kernel-based classifiers. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3815–3824, 2019.

Kenny Chen. A quantum/classical example oracle separation for making things up. *arXiv preprint arXiv:2608.11648*, 2026.

Evan Peters, João Caldeira, Alan Ho, Stefan Leichenauer, Masoud Mohseni, Hartmut Neven, Panagiotis Spentzouris, Doug Strain, and Gabriel N. Perdue. Machine learning of high dimensional data on a noisy quantum processor. *npj Quantum Information*, 7(1):161, 2021. doi: 10.1038/s41534-021-00498-9.

Erik M. Åsgrim and Stefano Markidis. Quantum kernels are spectral tensor networks. *arXiv preprint arXiv:2606.20402*, 2026.

Oded Goldreich and Leonid A. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 25–32, 1989. doi: 10.1145/73007.73010.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012. doi: 10.1214/ECP.v17-2079.

Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

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

Ian T. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Igor L. Markov and Yaoyun Shi. Simulating quantum computation by contracting tensor networks. *SIAM Journal on Computing*, 38(3):963–981, 2008. doi: 10.1137/050644756.

Johnnie Gray and Stefanos Kourtis. Hyper-optimized tensor network contraction. *Quantum*, 5:410, 2021.

Norbert Schuch, Michael M. Wolf, Frank Verstraete, and J. Ignacio Cirac. Computational complexity of projected entangled pair states. *Physical Review Letters*, 98(14):140506, 2007. doi: 10.1103/PhysRevLett.98.140506.

Jonas Haferkamp, Dominik Hangleiter, Jens Eisert, and Marek Gluza. Contracting projected entangled pair states is average-case hard. *Physical Review Research*, 2(1):013010, 2020. doi: 10.1103/PhysRevResearch.2.013010.

Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14):147902, 2003. doi: 10.1103/PhysRevLett.91.147902.

Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill. Provably efficient machine learning for quantum many-body problems. *Science*, 377(6613):eabk3333, 2022. doi: 10.1126/science.abk3333.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Harry Buhrman, Richard Cleve, John Watrous, and Ronald de Wolf. Quantum fingerprinting. *Physical Review Letters*, 87(16):167902, 2001. doi: 10.1103/PhysRevLett.87.167902.

Juan Carlos Garcia-Escartin and Pedro Chamorro-Posada. SWAP test and Hong–Ou–Mandel effect are equivalent. *Physical Review A*, 87(5):052330, 2013. doi: 10.1103/PhysRevA.87.052330.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020. doi: 10.1038/s41567-020-0932-7.

Andreas Elben, Steven T. Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. *Nature Reviews Physics*, 5(1):9–24, 2023. doi: 10.1038/s42254-022-00535-2.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988. doi: 10.1016/0024-3795(88)90223-6.

Andreas Elben, Benoît Vermersch, Rick van Bijnen, Christian Kokail, Tiff Brydges, Christine Maier, Manoj K. Joshi, Rainer Blatt, Christian F. Roos, and Peter Zoller. Cross-platform verification of intermediate scale quantum devices. *Physical Review Letters*, 124(1):010504, 2020. doi: 10.1103/PhysRevLett.124.010504.

Valentin Heyraud, Zejian Li, Zakari Denis, Alexandre Le Boité, and Cristiano Ciuti. Noisy quantum kernel machines. *Physical Review A*, 106(5):052421, 2022. doi: 10.1103/PhysRevA.106.052421.

Organization of the appendices. Appendix A fixes the notation used throughout. Appendices B to F contain the proofs, grouped by subject: interface structure and capacity; the controlled factorization, spectrum and evaluation; the computational and learning frontier; finite-rank learning and the sample–shot trade-off; and the finite-measurement and sharp-interface results, which include the prefix-span bound of Proposition 8. Appendix G specifies the experimental setup shared by every study, and Appendix H reports the evidence, organized by the claim each study tests rather than by experiment type.

A Background, Notation, and Related Work

This appendix collects additional notation and formulations used in the proofs and provides the extended related-work discussion summarized in Section 2.

A.1 Quantum and Operator Background

Notation. We let $[n] = \{1, \dots, n\}$ for any positive integer n. We write $\otimes$ for the tensor product, $X^\dagger$ for the conjugate transpose of an operator X, Tr for the trace, and Tr_A for the partial trace over register A. A quantum state vector is denoted by a ket $|\psi\rangle$, its dual by a bra $\langle\psi| = |\psi\rangle^\dagger$, and the corresponding rank-one projector by $|\psi\rangle\langle\psi|$. The identity operator on register A is denoted by I_A. The state space of register A is $\mathcal{H}_A$, and $\mathrm{HS}(\mathcal{H})$ denotes the space of linear operators on $\mathcal{H}$ equipped with the Hilbert–Schmidt inner product $\langle X, Y\rangle_{\mathrm{HS}} = \mathrm{Tr}[X^\dagger Y]$. Vectors carry the standard Euclidean norm $\|\cdot\|$, while $O(\cdot)$ hides absolute constants.

Registers and interface quantities. Registers A and B denote the encoding and readout registers, containing n_A and n_B qubits, respectively, with total qubit count $N = n_A + n_B$. For circuit interfaces, b is the observed cut-width witness of Section 3.1, $b^\star_{\mathcal{G}}$ is the minimal crossing count within a specified gate library and allowed local equivalences, and $C_{\mathrm{OS}} = \log_2 r$, where $r = \mathrm{SR}(U_{\mathrm{ent}})$ is the operator Schmidt rank across the cut. We use b_{eff} and b_{tgt} for the numbers of distinct control and target qubits among the crossing gates, and m_β for the number of crossings sharing target qubit β.

Controlled interfaces. For controlled interfaces, $\{P_s\}$ denotes the projectors onto the control-basis subspaces and $L = 2^{b_{\mathrm{eff}}}$ their number. The vector $p(x) \in \mathbb{R}^L$ contains the corresponding input-dependent probabilities, $Q \in \mathbb{R}^{m\times L}$ is the interface matrix with entries $Q_{is} = p_s(x_i)$, and G denotes the fixed readout Gram matrix.

A.2 Kernel-Learning Formulation

Projected quantum kernels. In the learning problem, P denotes the input distribution, m the number of labeled samples, and K the empirical Gram matrix. The population projected feature dimension D is defined in Eq. (1). The linear projected quantum kernel is

$$k_B(x, x') = \mathrm{Tr}[\rho_B(x)\rho_B(x')], \qquad K_{ij} = k_B(x_i, x_j). \tag{7}$$

Population operators. The corresponding population integral operator and feature covariance operator are

$$(L_{k_B} f)(x) = \int k_B(x, x') f(x')\, \mathrm{d}P(x'), \qquad \Sigma T = \mathbb{E}[\langle \rho_B(x), T\rangle_{\mathrm{HS}} \rho_B(x)]. \tag{8}$$

The nonzero spectra of these operators coincide under the standard feature-operator correspondence, allowing the spectral quantities used in the main analysis to be expressed directly through the projected representation.

Statistical notation. We use ν for the approximation-envelope exponent in Eq. (55) and Y for the response bound in Proposition 4. Additional quantities used locally in the statistical results are introduced where needed.

A.3 Extended Related Work

Quantum computing and learning applications. Quantum complexity theory formalizes the computational capabilities of quantum devices (Bernstein and Vazirani, 1997); quantum simulation provides a central application (Lloyd, 1996; Georgescu et al., 2014). Quantum algorithms address a range of computational tasks (Montanaro, 2016), including dissipative nonlinear differential equations under specified parameter and input-access conditions (Liu et al., 2021b). In cryptanalysis, concrete resource estimates clarify the requirements for factoring cryptographic-size integers on fault-tolerant quantum hardware (Gidney and Ekerå, 2021). For learning, quantum embeddings and approximation results motivate the use of quantum feature spaces (Lloyd et al., 2020; Goto et al., 2021). Examples include tailored quantum-kernel classifiers (Blank et al., 2020), classification in high-energy physics (Wu et al., 2021), kernel regression and differential-equation solvers (Paine et al., 2023), and trapped-ion classification and regression (Suzuki et al., 2024a).

Projected quantum kernels and statistical capacity. Quantum kernel methods encode data into quantum states and use quantum-state similarities as kernels for classical learning. Trainable embedding kernels and covariant kernels adapt the feature map to the learning problem (Hubregtsen

et al., 2022; Glick et al., 2024), while quantum Fisher kernels address vanishing similarities using a different kernel construction (Suzuki et al., 2024b). Projected quantum kernels construct such kernels from reduced quantum states or local quantum features (Huang et al., 2021; Gan et al., 2023). A complementary line of work studies which functions quantum kernels learn efficiently. The eigenspectrum and inductive bias of a quantum kernel determine how target functions are distributed across its learnable directions (Kübler et al., 2021), while target–kernel alignment relates spectral structure to generalization (Canatar et al., 2021). Kernel bandwidth can determine whether the spectrum remains sufficiently informative (Shaydulin and Wild, 2022), and recent double-descent analyses characterize how the selected quantum feature dimension affects interpolation and prediction (Kempkes et al., 2026). Quantum-kernel architectures can also be adapted directly to improve learning performance (Lei et al., 2024). Trace-induced quantum kernels provide a broader framework for kernels constructed from quantum-state overlaps and reduced representations (Gan et al., 2023). Nonlinear outer kernels can induce large or even infinite-dimensional RKHSs from finite-dimensional reduced representations, but still operate only on the information contained in the reduced state. Active quantum subspaces take a related statistical perspective by defining a sample-regularized dimension through measured observables and studying when additional quantum directions improve prediction (Bang et al., 2026). These works generally begin with a quantum representation or collection of measured features that has already been specified. Our structural analysis instead derives the representation reaching the readout from the circuit connecting the encoding and readout subsystems.

Circuit-level constraints on quantum representations. Generalization bounds for trainable quantum models can depend on trainable gate counts and other circuit-level complexity measures (Caro et al., 2022). Data encoding can become concentrated as circuit depth increases, reducing the distinguishability of encoded states (Li et al., 2022). For data re-uploading models, circuit structure can also constrain the function classes that are representable (Yu et al., 2022). These results constrain the circuit or encoded representation before a particular projected readout is analyzed. Our structural quantities instead characterize how many independent operator directions can be transmitted to a specified readout and which of them are activated by the data.

Classical accessibility and quantum–classical learning separations. A statistically low-dimensional quantum representation need not be computationally easy to evaluate. Classical-surrogate methods identify quantum learning models whose input–output behavior can be reproduced efficiently by classical models (Schreiber et al., 2023). Other work studies connections between trainability and classical simulability (Cerezo et al., 2025) and between trainability and dequantization in variational quantum learning models (Gil-Fuster et al., 2025). Quantum algorithms can also accelerate the training of linear and kernel-based classifiers under suitable quantum input-access assumptions (Li et al., 2019). A separate literature establishes quantum–classical separations in learning. Early work proved separations between quantum and classical PAC learning using quantum-example access for the quantum learner (Servedio and Gortler, 2004). More recent results distinguish end-to-end quantum learning advantages from the hardness of evaluating a learned hypothesis (Molteni et al., 2026), and show how broader computational quantum advantages can induce supervised-learning advantages under explicit conditions (Yamasaki et al., 2026). Concurrent work also uses hard-core predicates in related oracle learning settings (Chen, 2026). Our computational analysis instead asks whether the projected representation itself can be evaluated classically, and our learning separation gives both learners ordinary i.i.d. classical labeled samples.

Finite measurements in quantum kernels. Measurement cost imposes a restriction distinct from structural dimension. Limited measurements and device noise can reduce the practical effectiveness of quantum kernels in NISQ settings (Wang et al., 2021), while exponential concentration can make kernel values difficult to resolve with finite measurements (Thanasilp et al., 2024). Hardware experiments likewise illustrate how finite sampling and noise affect practical kernel learning (Peters et al., 2021). Our finite-measurement analysis starts from a structurally reachable representation and asks which of its directions remain resolvable under a finite preparation budget. For controlled interfaces, the exact kernel factorization further allows the required state-preparation cost to be analyzed directly.

Overall, prior work either constrains the quantum model before a readout is fixed or studies the statistical, computational, and measurement properties of an already specified representation. The data-to-readout interface connects these stages by characterizing the input-dependent operator span that reaches the chosen readout.

B Interface Structure and Capacity Proofs

We begin with two auxiliary lemmas that will be used repeatedly in the proofs below. We then prove Theorem 1 and corollaries 1 and 2.

B.1 Structural capacity at the interface

Lemma 1 (Equivalent readings of the reachable dimension)**.** *Let $\varphi(x) = \rho_B(x)$ and assume that $x \mapsto \varphi(x)$ is continuous on* $\mathrm{supp}(P)$*. With L_{k_B} and Σ as defined in Eq. (8),*

$$\mathrm{rank}(L_{k_B}) = \mathrm{rank}(\Sigma) = \dim \mathrm{span}\{\varphi(x) : x \in \mathrm{supp}(P)\} = D. \tag{9}$$

For any $x_1, \ldots, x_m \in \mathrm{supp}(P)$, the associated Gram matrix satisfies $\mathrm{rank}(K) \le \min\{m, D\}$.

Proof. Let $\mathcal{W} = \mathrm{span}\{\varphi(x) : x \in \mathrm{supp}(P)\}$. Define $S : L^2(P) \to \mathrm{HS}(\mathcal{H}_B)$ by $Sf = \mathbb{E}[f(x)\varphi(x)]$, whose adjoint is $(S^*T)(x) = \langle \varphi(x), T\rangle$. Then

$$(S^*Sf)(x) = \int \langle \varphi(x), \varphi(x')\rangle f(x')\, \mathrm{d}P(x') = (L_{k_B} f)(x), \qquad SS^* = \Sigma.$$

Hence S^*S and SS^* have the same nonzero eigenvalues and $\mathrm{rank}(L_{k_B}) = \mathrm{rank}(\Sigma) = \mathrm{rank}(S)$.

Moreover, $\Sigma T = \mathbb{E}[\langle \varphi(x), T\rangle \varphi(x)]$ lies in $\mathcal{W}$, so $\mathrm{range}(\Sigma) \subseteq \mathcal{W}$. Conversely, if $v \perp \mathrm{range}(\Sigma)$, then

$$0 = \langle v, \Sigma v\rangle = \mathbb{E}\big|\langle \varphi(x), v\rangle\big|^2.$$

Thus $\langle \varphi(x), v\rangle = 0$ almost surely. Continuity extends this equality to all of $\mathrm{supp}(P)$, so $v \perp \mathcal{W}$. Therefore $\mathrm{range}(\Sigma) = \mathcal{W}$ and its dimension is D.

Finally, K is the Gram matrix of m feature operators in the D-dimensional space $\mathcal{W}$, giving $\mathrm{rank}(K) \le \min\{m, D\}$. □

Lemma 2 (Operator Schmidt rank across the interface)**.** *Suppose a realization of U_{ent} contains crossing gates $e \in \mathcal{C}$ with operator Schmidt ranks d_e. Then*

$$r = \mathrm{SR}_{A:B}(U_{\mathrm{ent}}) \le \prod_{e \in \mathcal{C}} d_e. \tag{10}$$

In particular, $r \le \kappa^b$ if there are b crossings and $d_e \le \kappa$ for all of them.

Proof. View an operator on $A \otimes B$ as an element of $\mathrm{HS}(\mathcal{H}_A) \otimes \mathrm{HS}(\mathcal{H}_B)$ (Nielsen et al., 2003). Local gates have operator Schmidt rank one, while operator Schmidt rank is submultiplicative under composition: expanding U and V in product-operator decompositions gives $\mathrm{SR}(UV) \le \mathrm{SR}(U)\mathrm{SR}(V)$. Applying this inequality gate by gate leaves only the crossing gates in the product and yields Eq. (10). The final statement follows immediately. □

Proof of Theorem 1. By definition, $\rho_B(x) = \mathcal{N}_{A\to B}(\rho_A(x))$, so

$$\mathrm{span}\{\rho_B(x) : x \in \mathrm{supp}(P)\} = \mathcal{N}_{A\to B}(\mathcal{V}_A).$$

This gives $D = \mathrm{rank}(\mathcal{N}_{A\to B}|_{\mathcal{V}_A}) \le r_{\mathrm{tr}}$.

To bound the full transfer rank, write an operator Schmidt decomposition $U_{\mathrm{ent}} = \sum_{k=1}^{r} \sigma_k O_A^{(k)} \otimes O_B^{(k)}$. For any Hermitian X,

$$\mathcal{N}_{A\to B}(X) = \sum_{k,l=1}^{r} \sigma_k \sigma_l \,\mathrm{Tr}\Big[O_A^{(k)} X O_A^{(l)\dagger}\Big]\, O_B^{(k)} |0\rangle\langle 0|_B O_B^{(l)\dagger}. \tag{11}$$

Every output therefore belongs to the fixed span of at most r^2 operators. Because the family is closed under adjoint, restricting to Hermitian outputs does not increase this dimension, and hence $r_{\mathrm{tr}} \le r^2$. Combining this with Lemma 2 proves Eq. (2) and the crossing-gate bound. If $\mathcal{V}_A$ is the full Hermitian operator space on A, the restriction does not reduce the channel rank and $D = r_{\mathrm{tr}}$.

For completeness, fix a gate library $\mathcal{G}$ and let $b^\star_{\mathcal{G}}$ be the minimum number of crossings among exact realizations of U_{ent}, with local equivalences treated as free. If every allowed crossing has operator Schmidt rank at most $\kappa_{\mathcal{G}}$, then Lemma 2 gives $\log_{\kappa_{\mathcal{G}}} r \le b^\star_{\mathcal{G}} \le b$. For a rank-two crossing library this becomes $\log_2 r \le b^\star_{\mathcal{G}} \le b$ and, together with Eq. (2), $D \le 4^{b^\star_{\mathcal{G}}} \le 4^b$. □

Proof of corollary 1. The final unitary W_A drops out under the partial trace. Absorb V_A into the encoder and set $|\psi(x)\rangle = V_A U_A(x)|0\rangle_A$ and $|\theta\rangle = V_B|0\rangle_B$. Let S be the set of distinct controls. Because all controlled gates are diagonal with respect to the same computational-basis sectors on S, their product can be written as

$$\prod_{i=1}^{b} \mathrm{CU}_{a_i \to \beta_i} = \sum_{s \in \{0,1\}^S} P_s \otimes U_s, \tag{12}$$

where $\{P_s\}$ are mutually orthogonal projectors summing to I_A and U_s is the ordered product of target actions selected by sector s. The joint state before the final local unitaries is therefore

$$\sum_s P_s|\psi(x)\rangle \otimes U_s|\theta\rangle.$$

Tracing out A removes all terms with $s \neq s'$ because $P_s P_{s'} = 0$. After restoring W_B, one obtains

$$\rho_B(x) = \sum_s p_s(x)\tau_s, \qquad p_s(x) = \langle\psi(x)|P_s|\psi(x)\rangle, \qquad \tau_s = W_B U_s|\theta\rangle\langle\theta|U_s^\dagger W_B^\dagger.$$

Consequently every readout state lies in $\mathrm{span}\{\tau_s\}$ and $D \le \dim \mathrm{span}\{\tau_s\} = \mathrm{rank}(G) \le 2^{b_{\mathrm{eff}}}$, where G is the Hilbert–Schmidt Gram matrix of the branch states. Since $b_{\mathrm{eff}} \le b$, the dimension bound follows.

Substituting the mixture into the Hilbert–Schmidt kernel gives

$$k_B(x, x') = \sum_{s,s'} p_s(x) p_{s'}(x') \,\mathrm{Tr}(\tau_s \tau_{s'}) = p(x)^\top G p(x'),$$

which proves the factorization. □

Proof of corollary 2. Let $\sigma_C(x)$ denote the state on the separator and $\rho_R(x) = \mathcal{E}(\sigma_C(x))$ the retained readout state after the fixed suffix. Define $\mathcal{V}_C = \mathrm{span}\{\sigma_C(x) : x \in \mathrm{supp}(P)\}$. Since C contains q qubits, $\dim \mathcal{V}_C \le 4^q$. Linearity of the fixed suffix gives

$$\mathrm{span}\{\rho_R(x) : x \in \mathrm{supp}(P)\} = \mathcal{E}(\mathcal{V}_C),$$

so $D_R \le \mathrm{rank}(\mathcal{E}) \le 4^q$. If the separator is classical and selects among fixed states $\tau_1, \dots, \tau_L$, every output lies in their span and $D_R \le L$.

A valid causal separator must contain every quantum or memory wire at the cut that can later influence R, and the suffix after the cut must contain no input-dependent operation. Appending fixed ancillas, applying the fixed suffix, and tracing out all systems except R then defines the required input-independent CPTP map $\mathcal{E}$. A generic time slice need not satisfy this condition when data are re-uploaded later or an omitted memory wire subsequently reaches the readout. □

The separator also yields a spectral contraction statement that is useful independently of the dimension ceiling. Choose real orthonormal Hermitian bases on C and R, let $s_C(x)$ and $z_R(x)$ be the corresponding coordinate vectors, and let E represent $\mathcal{E}$, so $z_R(x) = E s_C(x)$. With $C_C = \mathbb{E}[s_C(x) s_C(x)^\top]$, the nonzero population-kernel spectrum on R equals that of

$$M_R = E C_C E^\top. \tag{13}$$

Indeed, if $(Tw)(x) = w^\top z_R(x)$, then $L_{k_R} = TT^*$ and $T^*T = M_R$. Thus the two operators have the same nonzero eigenvalues. Moreover, writing $\chi_{\mathcal{E}} = \|E\|_{2\to 2}$ and using $s_j(AB) \le \|A\| s_j(B)$ gives $\lambda_j(M_R) \le \chi^2_{\mathcal{E}} \lambda_j(C_C)$ and hence $d_{\mathrm{eff},R}(\gamma) \le d_{\mathrm{eff},C}(\gamma/\chi^2_{\mathcal{E}})$.

B.2 CONTROLLED-INTERFACE REFINEMENTS

For arbitrary single-target unitaries, define $\mathcal{A}_\beta = \{\prod_{i:\beta_i=\beta}^{\rightarrow} U_i^{s_{a_i}} : s \in \{0,1\}^{b_{\text{eff}}}\}$ and $L_\beta = |\mathcal{A}_\beta|$. Since actions on different targets commute, the number of global branch actions is at most $\prod_\beta L_\beta$, so

$$D \le \min\{2^{b_{\text{eff}}}, \textstyle\prod_\beta L_\beta\}. \tag{14}$$

Shared controls can only reduce the number of jointly attainable actions. If all crossings on a target β use a common unitary U_β and m_β crossings reach that target, then

$$L_\beta \le \min\{m_\beta + 1, \mathrm{ord}(U_\beta)\}, \qquad D \le \prod_\beta L_\beta, \tag{15}$$

where $\mathrm{ord}(U_\beta)$ is its projective order. In particular, a common involution contributes at most two branches per distinct target.

Contiguity in Eq. (3) is sufficient but not necessary. Suppose fixed intra-A gates are interleaved but commute with every P_s, while arbitrary fixed intra-B gates are allowed between crossings. Then the joint evolution maps each initial sector to $A_s P_s|\psi(x)\rangle \otimes V_s|\theta\rangle$, with A_s unitary inside sector s. The sector vectors remain orthogonal and retain squared norm $p_s(x)$, so tracing out A again gives

$$\rho_B(x) = \sum_s p_s(x) V_s|\theta\rangle\langle\theta|V_s^\dagger. \tag{16}$$

Thus the factorization and $2^{b_{\text{eff}}}$ ceiling remain valid. The target-wise refinements above need the stronger product-action structure and do not follow from arbitrary interleaved readout gates.

Remark 1 (Necessity of the target-side assumptions). The common-unitary and involution assumptions in the target-side refinements play distinct roles. For the involution condition, consider two distinct controls acting on one target through the same controlled $R_y(\alpha)$, with the target initialized in $|0\rangle$ and $\alpha \notin \pi\mathbb{Z}$. Computational-basis control inputs generate

$$\rho_k = R_y(k\alpha)|0\rangle\langle 0|R_y(k\alpha)^\dagger, \qquad k \in \{0,1,2\}.$$

In the Hermitian basis $\{I, X, Z\}$ their coefficient vectors are proportional to $(1,0,1)$, $(1, \sin\alpha, \cos\alpha)$, and $(1, \sin 2\alpha, \cos 2\alpha)$. The determinant of the resulting coefficient matrix is $2(\cos\alpha - 1)\sin\alpha \neq 0$, so $D = 3$ although there is only one target qubit. This agrees with the $m_\beta + 1$ count before the involution refinement.

The common-unitary condition is also necessary even if every target action is an involution. Let two controls act on the same target through controlled-X and controlled-Z, and choose $\rho = (I + (X + Y + Z)/\sqrt{3})/2$. The four control patterns produce ρ, $X\rho X$, $Z\rho Z$, and $Y\rho Y$. Their Hilbert–Schmidt Gram spectrum is $(2, 2/3, 2/3, 2/3)$, so the four states are linearly independent and $D = 4$ with one target. Thus the common-unitary premise cannot be removed from the target-wise refinement.

B.3 LOCALITY AND FUNCTION-SPACE TENSOR-NETWORK SEPARATION

Lemma 3 (Euclidean one-crossing light-cone bound)**.** *Suppose m raw binary features are loaded, without input-dependent preprocessing, onto m distinct sites of a fixed s-dimensional bounded-degree lattice, after which all circuit gates are input-independent. Assume all encoder gates before the unique A:B crossing have bounded range and bounded arity, and the encoder-side support of that crossing has $O(1)$ sites. If the final reduced readout state depends nontrivially on each of the m raw features, then the encoder depth before the crossing is $\Omega(m^{1/s})$.*

Proof. Let $Q \subset A$ be the encoder-side support of the unique crossing gate. Since no later A:B gate occurs, gates acting only on A after the crossing disappear under the partial trace and cannot alter the reduced state on B. Therefore all input dependence that reaches B must already be present in the pre-crossing reduced state on Q.

In a depth-t bounded-range circuit on a fixed s-dimensional lattice, the backward light cone of the constant-size set Q lies in a ball of radius $O(t)$ and contains at most $O(t^s)$ sites. A feature loaded outside this light cone cannot affect the state on Q and hence cannot affect the readout. Since every

one of the m raw features is essential, all loading sites must lie in the light cone. Thus $m = O(t^s)$ and $t = \Omega(m^{1/s})$. □

Proposition 1 (Physical-interface and function-space tensor-network complexities are incomparable)**.** *Fix the ordering of d binary input-feature sites, and let $\chi_{\mathrm{func}}(k)$ be the minimum exact MPO bond dimension of the coefficient tensor of a kernel $k(x, x')$ in a complete local product basis over those sites. Then the physical-interface quantities $\log_2 r$, $b^\star_{\mathcal{G}}$, and D are incomparable with χ_{func}: there are one-crossing families with $D = 2$ and $\chi_{\mathrm{func}} = 2^d$, and d-crossing families with $D = 2^d$ and $\chi_{\mathrm{func}} = 1$.*

Proof. We give explicit constructions in both directions. Exact MPO bond dimension across a feature-site cut equals the matrix rank of the corresponding coefficient-tensor matricization, as in the standard tensor-train/MPO rank characterization (Åsgrim and Markidis, 2026).

For the first family, let $d = 2n$, write $x = (u, v)$ with $u, v \in \{0, 1\}^n$, and define $g(x) = (-1)^{u\cdot v}$ and $c(x) = (1 - g(x))/2$. The encoder computes $c(x) = \bigoplus_i u_i v_i$ using only local gates on A, then a single CNOT copies this bit to a blank readout qubit. Thus

$$\rho_B(x) = |c(x)\rangle\langle c(x)|, \qquad k(x, x') = \frac{1 + g(x)g(x')}{2}. \tag{17}$$

The interface has $r = 2$, $b^\star_{\mathcal{G}} = 1$ for a rank-two library containing CNOT, and $D = 2$.

The encoder computation can be implemented with $O(d)$ bounded-arity Clifford+Toffoli gates and depth $O(\log d)$ on a bounded-degree binary-tree interaction graph. Under strict s-dimensional bounded-range locality with local basis loading, place the input gadgets at constant density and accumulate the parity toward a designated root. This gives $O(d)$ gates and depth $O(d^{1/s})$. Since every raw bit is essential for $c(x)$, Lemma 3 supplies the matching lower bound $\Omega(d^{1/s})$.

It remains to determine the exact function-space bond dimension. Group the first n paired sites (u, u') against the last n paired sites (v, v'), and let $H_{u,v} = (-1)^{u\cdot v}$ be the $2^n \times 2^n$ Walsh–Hadamard matrix. Up to the factor $1/2$, the matricization is $J + H \otimes H$, where $J = \mathbf{1}\mathbf{1}^\top$. Setting $A = H \otimes H$, the identities $H^2 = 2^n I$, $A^{-1} = 2^{-2n} A$, and $A\mathbf{1} = 2^{2n} e_{00}$ give $\mathbf{1}^\top A^{-1}\mathbf{1} = 1$. The matrix determinant lemma therefore yields

$$\det(A + J) = 2\det(A) \neq 0.$$

The cut rank is consequently $2^{2n} = 2^d$, so $\chi_{\mathrm{func}}(k) = 2^d$ for the stated ordering. Any site-local invertible basis change, including the local Walsh basis, preserves this cut rank.

For the converse family, let $x \in \{0, 1\}^d$, prepare $|x\rangle_A$, and apply d disjoint CNOTs from A_j to blank readout qubits B_j. Then

$$\rho_B(x) = |x\rangle\langle x|, \qquad k(x, x') = \mathbf{1}\{x = x'\} = \prod_{j=1}^{d} \mathbf{1}\{x_j = x'_j\}.$$

The kernel is an exact product and therefore has $\chi_{\mathrm{func}} = 1$ for every ordering, whereas the 2^d orthogonal readout projectors give $D = 2^d$. The entangler has $r = 2^d$, and for a rank-two crossing library containing CNOT, Lemma 2 together with the explicit circuit gives $b^\star_{\mathcal{G}} = d$.

These two constructions establish both directions of incomparability. □

C Controlled Factorization and Effective-Capacity Proofs

This appendix derives the statistical and computational quantities that follow from the controlled-interface factorization in corollary 1. We first prove the population spectrum, target approximation, and nested-widening statements in corollary 3, and then give the coordinate-free nesting result and finite-dimensional evaluator used later in the paper.

C.1 SPECTRUM AND TARGET UTILITY

Proof of corollary 3. Let $L = 2^{b_{\text{eff}}}$, $z(x) = G^{1/2}p(x) \in \mathbb{R}^L$, and define $T : \mathbb{R}^L \to L^2(P)$ by $(Tw)(x) = w^\top z(x)$. Its adjoint is $T^* f = \mathbb{E}[z(x)f(x)]$. By Eq. (4),

$$L_{k_B} = TT^*, \qquad T^*T = \mathbb{E}[z(x)z(x)^\top] = G^{1/2}C_pG^{1/2} = M. \tag{18}$$

The operators TT^* and T^*T have the same nonzero eigenvalues, including multiplicities. Hence the nonzero population-kernel spectrum is exactly the nonzero spectrum of M. Under the continuity condition of Lemma 1, this also gives $D = \operatorname{rank}(M)$. The effective-dimension identity in Eq. (5) follows by applying $t \mapsto t/(t+\gamma)$ to these eigenvalues.

The RKHS of the linear feature map z consists of functions $h_w(x) = w^\top z(x)$, modulo null directions. With $c = \mathbb{E}[z(x)f^\circ(x)]$,

$$\mathbb{E}[(h_w(x) - f^\circ(x))^2] = w^\top Mw - 2w^\top c + \mathbb{E}[f^\circ(x)^2].$$

If $v \in \ker M$, then $\mathbb{E}[(v^\top z(x))^2] = 0$, so $v^\top c = 0$ and therefore $c \in \operatorname{range}(M)$. The minimum-norm optimizer is $w^\star = M^\dagger c$, and substitution yields the approximation formula in Eq. (5).

It remains to characterize nested enlargement under ridge regularization. For the current feature z, set $A_\gamma = M + \gamma I$. The regularized population objective has minimum

$$\mathbb{E}[f^\circ(x)^2] - c^\top A_\gamma^{-1} c. \tag{19}$$

Now append features $u(x)$, and write $C_{zu} = \mathbb{E}[zu^\top]$, $C_u = \mathbb{E}[uu^\top]$, and $c_u = \mathbb{E}[uf^\circ]$. The widened Hessian is

$$H_\gamma = \begin{pmatrix} A_\gamma & C_{zu} \\ C_{zu}^\top & C_u + \gamma I \end{pmatrix}, \qquad q = \begin{pmatrix} c \\ c_u \end{pmatrix}.$$

Its Schur complement $S_\gamma = C_u + \gamma I - C_{zu}^\top A_\gamma^{-1} C_{zu}$ is positive definite. The block inverse identity gives

$$q^\top H_\gamma^{-1} q = c^\top A_\gamma^{-1} c + r_\gamma^\top S_\gamma^{-1} r_\gamma, \qquad r_\gamma = c_u - C_{zu}^\top A_\gamma^{-1} c.$$

Comparing with Eq. (19), the widened optimum is strictly smaller exactly when $r_\gamma \neq 0$. □

For an unregularized nested pair $\mathcal{H}_b \subseteq \mathcal{H}_{b+1}$, the same conclusion has the familiar projection form. If A_b denotes the best population squared-loss approximation error in $\mathcal{H}_b$, then

$$A_b - A_{b+1} = \|\Pi_{\mathcal{H}_{b+1}} f^\circ - \Pi_{\mathcal{H}_b} f^\circ\|^2_{L^2(P)}. \tag{20}$$

Thus widening helps exactly when the larger space contains a target component absent from the smaller one.

C.2 NESTED INTERFACE MAPS AND REGULARIZATION GEOMETRY

Proposition 2 (Regularizer-compatible nested interface maps)**.** *Let a wider feature be $\widetilde{z}(x) \in \mathbb{R}^{d'}$ and suppose the narrower feature satisfies $z(x) = T\widetilde{z}(x)$ for a full-row-rank matrix $T \in \mathbb{R}^{d \times d'}$. Let $\mathcal{S} = \operatorname{range}(T^\top)$ and $P_{\mathcal{S}}$ be the Euclidean projector onto $\mathcal{S}$. Compare the wider Euclidean-ridge objective with the narrower objective using the pullback penalty $\gamma\|T^\top w\|_2^2$. With $\widetilde{M} = \mathbb{E}[\widetilde{z}\widetilde{z}^\top]$, $\widetilde{c} = \mathbb{E}[\widetilde{z}f^\circ]$, and $a_{\mathcal{S},\gamma}$ the restricted wider-space ridge minimizer over $\mathcal{S}$, the wider optimum is strictly smaller if and only if*

$$(I - P_{\mathcal{S}})\big(\widetilde{c} - \widetilde{M}a_{\mathcal{S},\gamma}\big) \neq 0. \tag{21}$$

If $TT^\top = I_d$, the pullback penalty is the ordinary Euclidean ridge penalty in the narrow coordinates.

Proof. Because T has full row rank, $w \mapsto a = T^\top w$ is a bijection from $\mathbb{R}^d$ onto $\mathcal{S}$. Moreover, $w^\top z(x) = a^\top \widetilde{z}(x)$ and $\|T^\top w\|_2 = \|a\|_2$. Hence the narrow objective with the pullback penalty is exactly the wider Euclidean-ridge objective restricted to $\mathcal{S}$.

Choose orthonormal matrices U and V spanning $\mathcal{S}$ and $\mathcal{S}^{\perp}$, respectively. In coordinates $Q^{\top}\widetilde{z} = (U^{\top}\widetilde{z}, V^{\top}\widetilde{z}) =: (z_{\mathcal{S}}, u)$ with $Q = [U\ V]$, the restricted problem uses only $z_{\mathcal{S}}$ and the full problem appends u. The result just proved for corollary 3 therefore applies. Writing $a_{\mathcal{S},\gamma} = U\alpha_{\gamma}$, the correlation of the restricted ridge residual with the added coordinates is

$$V^{\top}(\widetilde{c} - \widetilde{M}a_{\mathcal{S},\gamma}).$$

The full optimum is strictly smaller exactly when this vector is nonzero. Since $VV^{\top} = I - P_{\mathcal{S}}$, this is equivalent to Eq. (21). If $TT^{\top} = I_d$, then $\|T^{\top}w\|_2 = \|w\|_2$, giving the final statement. □

C.3 Interface evaluator

Proposition 3 (Interface evaluator and computational localization)**.** *Under the controlled factorization with $L = 2^{b_{\text{eff}}}$ and $Q_{is} = p_s(x_i)$, the Gram matrix is $K = QGQ^{\top}$. Given Q and G, direct assembly costs $O(mL^2 + m^2L)$ with $O(L^2 + mL)$ working memory; explicit square-root features add $O(L^3)$. If $p(x)$ is classically computable to inverse-polynomial ℓ_1 precision and G to inverse-polynomial entrywise precision in polynomial time, and L is polynomial in the problem size, the downstream kernel construction is classically polynomial-time.*

Proof of Proposition 3. Applying Eq. (4) entrywise gives $K = QGQ^{\top}$. Forming QG costs $O(mL^2)$ and multiplying by $Q^{\top}$ costs $O(m^2L)$, with the stated memory. Since $G \succeq 0$, an explicit square root $G = RR^{\top}$ costs $O(L^3)$ and gives features $Z = QR$ with $K = ZZ^{\top}$.

Suppose $\hat{p}_i$ are probability vectors satisfying $\|\hat{p}_i - p_i\|_1 \le \varepsilon$ and $\hat{G}$ is clipped entrywise to $[0, 1]$ with $\|\hat{G} - G\|_{\max} \le \varepsilon_G$. Expanding the bilinear form gives

$$|\hat{p}_i^{\top}\hat{G}\hat{p}_j - p_i^{\top}Gp_j| \le |(\hat{p}_i - p_i)^{\top}\hat{G}\hat{p}_j| + |p_i^{\top}\hat{G}(\hat{p}_j - p_j)| + |p_i^{\top}(\hat{G} - G)p_j| \tag{22}$$

$$\le 2\varepsilon + \varepsilon_G, \tag{23}$$

so

$$\|\hat{Q}\hat{G}\hat{Q}^{\top} - QGQ^{\top}\|_{\max} \le 2\varepsilon + \varepsilon_G. \tag{24}$$

Measuring the controls samples $p(x)$ directly, while the fixed branch overlaps can be estimated once and reused across inputs. If both quantities are instead available classically to inverse-polynomial accuracy in polynomial time and L is polynomial, Eq. (24) shows that the entire kernel matrix can be assembled classically to inverse-polynomial precision in polynomial time.

When a positive-semidefinite estimate is required, project a symmetric estimate $\hat{G}$ onto the PSD cone and call the result G_+. Since $G \succeq 0$, $\|\hat{G} - G\|_2 \le L\varepsilon_G$ implies $\|G_+ - G\|_2 \le 2L\varepsilon_G$. Repeating the preceding expansion against the true G yields

$$\|\hat{Q}G_+\hat{Q}^{\top} - K\|_{\max} \le 2\varepsilon + 2L\varepsilon_G. \tag{25}$$

If G is exact, the empirical-probability estimator is already PSD and has rank at most L. □

D Computational and Learning Frontier Proofs

This appendix proves the computational frontier in two steps. The first localizes hardness in either the input-dependent probabilities or the reusable branch geometry. The second establishes the end-to-end learning witness and then transfers established learning separations to narrow projected interfaces.

D.1 Evaluation Hardness at Constant Statistical Capacity

Lemma 4 (Hard interface probabilities at constant capacity)**.** *Assume* BQP $\not\subseteq$ BPP. *There exists a uniform family of controlled-interface linear PQKs with a single crossing ($b = b_{\text{eff}} = 1$), orthogonal branch geometry $G = I$, and population feature dimension $D \le 2$ such that a designated kernel entry can be estimated on a quantum computer to additive error ε using $O(\varepsilon^{-2}\log(1/\delta))$ circuit executions, whereas a probabilistic classical polynomial-time algorithm that estimates that entry to*

any fixed additive error below $1/6$ *with success probability at least* $2/3$ *for every instance would imply* $\mathrm{BQP} \subseteq \mathrm{BPP}$. *The same two-dimensional coordinate linearly separates the yes/no promise classes with constant margin.*

Lemma 5 (Hard branch geometry with trivial probabilities)**.** *Assume* $\mathrm{BQP} \not\subseteq \mathrm{BPP}$. *There exists a family of two-branch classical interfaces in the sense of corollary 2 with deterministic, classically trivial branch probabilities and* $D \leq 2$ *such that estimating one reusable branch-overlap entry* G_{01} *to a fixed additive error below* $1/12$ *(any threshold below* $1/6$ *suffices) for every instance is BQP-hard for probabilistic classical polynomial time, while a quantum device estimates the same overlap with* $O(\varepsilon^{-2}\log(1/\delta))$ *branch-state preparations.*

Proof of Lemma 4. Let $\mathcal{L}$ be any language in BQP and let U_x be its uniform polynomial-size circuit family, with a designated acceptance qubit whose acceptance probability is $a(x) = \Pr[1 \mid U_x]$. By the bounded-error definition of BQP, we may take $a(x) \geq 2/3$ for yes instances and $a(x) \leq 1/3$ for no instances (Bernstein and Vazirani, 1997). Use U_x as the data-dependent encoder on register A, append one blank readout qubit B, and apply a single CNOT from the acceptance qubit to B. After tracing out A,

$$\rho_B(x) = (1 - a(x))|0\rangle\langle 0| + a(x)|1\rangle\langle 1|. \tag{26}$$

This architecture has $b = b_{\mathrm{eff}} = 1$ and hence $D \leq 2$ by corollary 1. Add one fixed reference input $x_\star$ whose encoder prepares the control qubit deterministically in $|1\rangle$, so that $\rho_B(x_\star) = |1\rangle\langle 1|$. The linear projected kernel then satisfies

$$k_B(x, x_\star) = \mathrm{Tr}[\rho_B(x)|1\rangle\langle 1|] = a(x). \tag{27}$$

A quantum device estimates this entry by repeated computational-basis measurements of B; Hoeffding's inequality gives $O(\varepsilon^{-2}\log(1/\delta))$ circuit executions for additive error ε and failure probability δ.

Suppose instead that a probabilistic classical polynomial-time algorithm estimated Eq. (27) to additive error strictly smaller than $1/6$ with success probability at least $2/3$ for every x. Thresholding the estimate at $1/2$ would decide $\mathcal{L}$ in BPP, because the yes and no cases are separated by the interval $[1/3, 2/3]$. Since $\mathcal{L}$ was arbitrary, such an estimator would imply $\mathrm{BQP} \subseteq \mathrm{BPP}$. Under the stated assumption, no such classical estimator exists for all instances. Finally, the two interface coordinates are $p_0(x) = 1 - a(x)$ and $p_1(x) = a(x)$. The linear function $h(x) = p_1(x) - p_0(x) = 2a(x) - 1$ has margin at least $1/3$ on the yes/no promise classes, so the same constant-dimensional representation is statistically simple while its evaluation remains computationally hard classically. □

Proof of Lemma 5. Let $\mathcal{L}$ be any language in BQP and let w denote an instance. Let V_w be its uniform verifier circuit on a work register W, with acceptance projector Π_1 on a designated output qubit and acceptance probability

$$a(w) = \langle 0|V_w^\dagger \Pi_1 V_w|0\rangle.$$

Append one ancilla qubit a initialized to $|0\rangle$, apply V_w on W, copy the acceptance qubit to a by CNOT, and uncompute V_w on W. Denote the resulting pure state by

$$|\psi_w\rangle = V_w^\dagger \,\mathrm{CNOT}_{\mathrm{out}\to a}\, V_w|0\rangle_W|0\rangle_a.$$

For the kernel instance indexed by w, let the two fixed branch states be $\tau_0^{(w)} = |0_W, 1_a\rangle\langle 0_W, 1_a|$ and $\tau_1^{(w)} = |\psi_w\rangle\langle\psi_w|$. Their Hilbert–Schmidt overlap is

$$G_{01}^{(w)} = \mathrm{Tr}(\tau_0^{(w)}\tau_1^{(w)}) = |\langle 0_W, 1_a|\psi_w\rangle|^2 = a(w)^2, \tag{28}$$

because projecting the copied ancilla onto 1 inserts Π_1 between V_w and $V_w^\dagger$. Within each such kernel instance, consider a two-point data domain $t \in \{0, 1\}$ whose classical interface probabilities are deterministic, $p(t) = e_t$. Then the projected states are exactly $\tau_t^{(w)}$, so $D \leq 2$, the probabilities are classically trivial, and the cross-kernel entry equals $G_{01}^{(w)}$. For yes instances $a(w) \geq 2/3$ and hence $G_{01}^{(w)} \geq 4/9$; for no instances $a(w) \leq 1/3$ and hence $G_{01}^{(w)} \leq 1/9$. Therefore any probabilistic classical polynomial-time procedure that estimates G_{01} to additive error below $1/12$ with success probability at least $2/3$ would decide $\mathcal{L}$ in BPP by thresholding between these intervals,

contradicting BQP$\not\subseteq$BPP. A quantum device can prepare $|\psi_w\rangle$ and estimate the probability of the basis outcome $|0_W, 1_a\rangle$ directly; Hoeffding concentration gives $O(\varepsilon^{-2}\log(1/\delta))$ preparations for additive error ε. This proves the claim. □

Proof of the evaluation clauses in Theorem 2. The easy direction is exactly Proposition 3: with polynomially many interface coordinates, inverse-polynomial ℓ_1 access to $p(x)$ and entrywise access to G gives polynomial-time assembly of K to inverse-polynomial max-norm precision. Any downstream kernel algorithm whose running time and required arithmetic precision are polynomial in the problem description can therefore be executed classically from this approximation. For the first hard direction, Lemma 4 has two orthogonal branch states, so $G = I$ and all hardness resides in the probability vector $p(x)$ while $D \le 2$. For the complementary direction, Lemma 5 has deterministic probabilities and $D \le 2$, while all hardness resides in one reusable overlap entry of G. Thus constant statistical capacity is compatible with both classically easy and classically hard interface realizations, proving that the listed capacity quantities cannot alone certify classical accessibility. □

D.2 End-to-End Learning Hardness and Inherited Realizations

Lemma 6 (Low-output feature maps admit narrow-interface PAC realizations)**.** *For any deterministic feature map $f_N : \{0,1\}^N \to \mathbb{F}_2^{d_N}$, there is a controlled-interface linear projected representation using d_N crossing CNOTs and d_N readout qubits such that*

$$\rho_B(x) = |f_N(x)\rangle\langle f_N(x)|, \qquad k_B(x,x') = \mathbf{1}\{f_N(x) = f_N(x')\}. \tag{29}$$

Writing $\pi_z = \Pr[f_N(X) = z]$, the nonzero population-kernel eigenvalues are exactly $\{\pi_z : \pi_z > 0\}$, so

$$D = |\operatorname{supp}(f_N(X))| \le 2^{d_N}, \qquad d_{\text{eff}}(\gamma) = \sum_{z:\pi_z>0} \frac{\pi_z}{\pi_z + \gamma}. \tag{30}$$

Every signed parity concept $(-1)^{f_N(x)\cdot s}$ is an exact linear functional of the readout state. If f_N satisfies the quantum-advantage conditions of Yamasaki et al. (2026, Theorem 1), their quantum PAC learnability and classical hardness transfer to this projected representation. In particular, $d_N = O(\log N)$ gives logarithmic interface width and polynomial D, d_{eff}.

Corollary 5 (Constant-dimensional supervised learning witness)**.** *Assume the classical hardness of discrete logarithm used by Liu et al. (2021a). There exists a family of linear projected quantum kernels with a one-qubit data-to-readout separator and population feature dimension $D = 3$ for which their discrete-log concept class is efficiently learnable by a quantum learner, whereas no probabilistic classical polynomial-time learner achieves test accuracy $1/2 + 1/\operatorname{poly}(n)$ under the uniform input distribution. Every target concept is exactly realizable by a linear functional of the projected readout state.*

Lemma 7 (Observable-to-interface compression)**.** *For M efficiently measurable two-outcome observables $\{P_i\}_{i=1}^M$ with $\mu_i(x) = \operatorname{Tr}[\rho_U(x)P_i]$, there exists a controlled interface with at most $2M$ populated orthogonal branches, $b_{\text{eff}} = \lceil \log_2(2M) \rceil$, $G = I$, and probabilities*

$$p_{i,\pm}(x) = \frac{1 \pm \mu_i(x)}{2M}. \tag{31}$$

Hence $D \le 2M$, and every target $f_\alpha(x) = \sum_i \alpha_i \mu_i(x)$ is represented exactly as

$$f_\alpha(x) = M \sum_{i=1}^{M} \alpha_i \big(p_{i,+}(x) - p_{i,-}(x)\big). \tag{32}$$

If $M = \operatorname{poly}(n)$, the interface width is $O(\log n)$ while the whole target family remains in the projected linear span.

Corollary 6 (Logarithmic-interface identification witness)**.** *Under the complexity assumption and hard observable family of Molteni et al. (2026, Theorem 15), there exists a controlled-interface projected representation with $b_{\text{eff}} = O(\log n)$ and $D = \operatorname{poly}(n)$ whose target family is quantumly identifiable but not classically identifiable by the approximate-correct learners considered there.*

Proof of Lemma 6. Let F be a d_N-qubit register on the encoder side. For the ideal deterministic feature map, absorb into the encoder a reversible computation that prepares $|f_N(x)\rangle_F$ (with any workspace uncomputed before the interface). Append a blank d_N-qubit readout B and apply one CNOT $F_j \to B_j$ for each feature bit j. This is a contiguous controlled interface with $b = b_{\text{eff}} = d_N$, and after tracing out the encoder,

$$\rho_B(x) = |f_N(x)\rangle\langle f_N(x)|. \tag{33}$$

The branch projectors are orthonormal in the Hilbert–Schmidt inner product, so $G = I$ and

$$k_B(x, x') = \mathrm{Tr}[|f_N(x)\rangle\langle f_N(x)||f_N(x')\rangle\langle f_N(x')|] = \mathbf{1}\{f_N(x) = f_N(x')\}. \tag{34}$$

Let $\mathcal{Z} = \{z : \pi_z > 0\}$ with $\pi_z = \Pr[f_N(X) = z]$. The normalized fiber indicators

$$u_z(x) = \frac{\mathbf{1}\{f_N(x) = z\}}{\sqrt{\pi_z}}, \qquad z \in \mathcal{Z}, \tag{35}$$

form an orthonormal set in $L^2(P)$. Direct substitution into the kernel integral operator gives $L_{k_B} u_z = \pi_z u_z$. Functions orthogonal to all fiber indicators are annihilated by L_{k_B}, proving that the complete nonzero spectrum is $\{\pi_z : z \in \mathcal{Z}\}$. Therefore $D = |\mathcal{Z}| \le 2^{d_N}$ and the effective-dimension identity in Eq. (30) follows immediately.

For $s \in \mathbb{F}_2^{d_N}$ define the diagonal Hermitian observable

$$W_s = \sum_{z \in \mathbb{F}_2^{d_N}} (-1)^{z \cdot s} |z\rangle\langle z|. \tag{36}$$

Then

$$\mathrm{Tr}[\rho_B(x) W_s] = (-1)^{f_N(x) \cdot s} = (-1)^{c_s(x)}, \tag{37}$$

so every concept is exactly represented by a linear functional of the projected state. Moreover $\|W_s\|_{\mathrm{HS}} = 2^{d_N/2}$, which is polynomial whenever $d_N = O(\log N)$.

It remains to connect the ideal representation to the computational learning statement. Yamasaki et al. (2026, Theorem 1) consider exactly the concept class $c_s(x) = f_N(x) \cdot s$ and prove quantum efficient PAC learnability together with the nonexistence of a classical efficient learner with an efficiently evaluatable hypothesis class under their stated condition that f_N is quantumly advantageous relative to randomized polynomial advice. Their quantum learner may compute $f_N(x)$ only with high probability on the target distribution; amplification makes the feature-computation failure probability inverse-polynomially smaller than the requested PAC tolerance, and their theorem already accounts for this distributional approximation. The controlled circuit above is therefore used only to characterize the ideal feature representation and its interface width; it is not asserted to be the exact circuit implementation used by their learner. Since the cited theorem concerns the same concept class and distributional feature map, its end-to-end quantum learner and classical lower bound provide the inherited learning statement used in the corollary. If $d_N = O(\log N)$, then $b = d_N = O(\log N)$ and $D, d_{\text{eff}}(\gamma) \le 2^{d_N} = \mathrm{poly}(N)$ for every $\gamma > 0$. □

Lemma 8 (Hard-core interface dictionary)**.** *Let $F_n : \{0,1\}^n \to \mathcal{Y}_n$ be an injective polynomial-time computable family that is classically one-way on a uniformly random preimage and quantum-polynomial-time invertible on its image, as assumed in the learning clause of Theorem 2. Set $L = 2^{\lfloor \log_2 n \rfloor}$. Let $\mathcal{R}_L$ be the uniform distribution over ordered linearly independent L-tuples R of masks in $\mathbb{F}_2^n$, and for $R \sim \mathcal{R}_L$ and $t \sim \mathrm{Unif}[L]$ let $f_{R,t}$ be defined as in the proof below. For a learner $\mathcal{A}$ that receives the public R and an i.i.d. labeled sample S_m of size m, define its prediction advantage*

$$\mathrm{Adv}(\mathcal{A}, m, n) = 2 \Pr[\mathcal{A}(R, S_m)(y^\star) = f_{R,t}(y^\star)] - 1,$$

the probability being over $R \sim \mathcal{R}_L$, t, the sample S_m, a fresh test point $y^\star$, and the internal randomness of $\mathcal{A}$. Then:

1. *the construction is a controlled interface with $b = b_{\text{eff}} = \log_2 L + 1$, $G = I_{2L}$ and $D = L + 1$, with spectrum Eq. (39);*
2. *a quantum learner achieves $\mathrm{Adv} \ge 1 - 2\delta$ from $m = \lceil \log_2((L-1)/\delta) \rceil$ ordinary labeled examples;*

3. *for every probabilistic polynomial-time* $\mathcal{A}$ *and every* $m = \mathrm{poly}(n)$, $\mathrm{Adv}(\mathcal{A}, m, n) = n^{-\omega(1)}$.

Proof of Lemma 8, and of the learning clause in Theorem 2. We first construct the interface family and compute its spectrum. Let $F_n : \{0,1\}^n \to \mathcal{Y}_n$ be injective, polynomial-time computable, classically one-way for a uniformly random preimage, and quantum-polynomial-time invertible on its image. Set $L = 2^{\lfloor \log_2 n \rfloor}$ and let $R = (r_1, \ldots, r_L)$ be a public ordered family of linearly independent masks in $\mathbb{F}_2^n$. For $y = F_n(a)$ define

$$h_j(y) = a \cdot r_j \pmod 2, \qquad f_{R,t}(y) = (-1)^{h_t(y)}, \quad t \in [L].$$

The exact representation below is defined by the unique preimage $a = F_n^{-1}(y)$; a bounded-error quantum inverter can be amplified to make its implementation error negligible. Exactness of $D = L + 1$ refers to this ideal representation; for the amplified circuit the crossing count still gives $D \le 2^b = 2L$ unconditionally, so the $\Theta(n)$ scaling in Theorem 2 is unaffected. A quantum encoder recovers a, prepares a uniform index register $L^{-1/2} \sum_{j=1}^{L} |j\rangle$, and coherently computes $h_j(y)$ into one qubit. Because the public dictionary R contains only $Ln = O(n^2)$ bits, the indexed parity computation can be compiled into a polynomial-size reversible circuit; no oracle access to an exponentially large table is being assumed. Copying the $\log_2 L$ index qubits and this predicate qubit to a blank readout uses $b = \log_2 L + 1$ crossing CNOTs. Tracing out the encoder gives

$$\rho_{B,R}(y) = \frac{1}{L} \sum_{j=1}^{L} |j, h_j(y)\rangle\langle j, h_j(y)|. \tag{38}$$

Thus the controlled branches are orthogonal, $G = I_{2L}$, and

$$p_R(y)_{j,c} = \frac{1}{L} \mathbf{1}\{c = h_j(y)\}.$$

Since R has full row rank and a is uniform, $(h_1(y), \ldots, h_L(y))$ is uniform on $\mathbb{F}_2^L$. Writing $d_j = e_{j,0} - e_{j,1}$ and $\mu = \mathbf{1}_{2L}/(2L)$,

$$p_R(y) = \mu + \frac{1}{2L} \sum_{j=1}^{L} (-1)^{h_j(y)} d_j.$$

Hence $M = C_p$ has one eigenvalue $1/(2L)$ in the constant direction and L eigenvalues $1/(2L^2)$ in the mutually orthogonal d_j directions. Therefore

$$D = L + 1, \qquad d_{\mathrm{eff}}(\gamma) = \frac{1}{1 + 2L\gamma} + \frac{L}{1 + 2L^2\gamma}. \tag{39}$$

Every target is exactly linear in the projected state:

$$f_{R,t}(y) = L \, \mathrm{Tr}\big[\big(|t,0\rangle\langle t,0| - |t,1\rangle\langle t,1|\big)\rho_{B,R}(y)\big],$$

so its approximation error is zero.

We next show that the quantum learner identifies the unknown target from labeled examples. The target index is genuinely learned from examples. Given m labeled examples $(y_i, f_{R,t}(y_i))$, a quantum learner inverts each y_i to a_i and compares the labels with the L candidate parities $a_i \cdot r_j$. For every $j \neq t$, the nonzero mask $r_j + r_t$ agrees on a uniformly random a_i with probability $1/2$. A union bound therefore gives

$$\Pr[\text{some } j \neq t \text{ remains consistent}] \le (L-1)2^{-m},$$

so $m \ge \lceil \log_2((L-1)/\delta) \rceil$ identifies t with probability at least $1 - \delta$. Prediction then inverts a fresh y quantumly and evaluates $a \cdot r_t$.

It remains to exclude a classical learner by reducing it to hard-core prediction. Suppose a probabilistic polynomial-time learner obtained non-negligible advantage over random guessing, averaged over a uniformly random ordered full-row-rank R, a uniformly random target index $t \in [L]$, its i.i.d. training sample, and a fresh test point. We use it to predict the Goldreich–Levin hard-core predicate

of F_n. Given a hard-core challenge $(F_n(a^\star), r^\star)$ with uniform $a^\star, r^\star \in \mathbb{F}_2^n$, the negligible event $r^\star = 0$ may be answered arbitrarily. Otherwise choose t uniformly and extend $r^\star$ to a uniformly random ordered linearly independent family R conditioned on $r_t = r^\star$; this can be done efficiently by sampling the remaining rows sequentially outside the span of those already chosen. The reduction can generate the learner's training set exactly: sample a_i uniformly, compute $y_i = F_n(a_i)$ in classical polynomial time, and attach the label $(-1)^{a_i \cdot r^\star}$. Run the learner on this public R and training set, and use its hypothesis on $F_n(a^\star)$ to predict $a^\star \cdot r^\star$. Conditioned on $r^\star \neq 0$, symmetry of the uniform full-row-rank ensemble implies that the resulting pair (R, t) has exactly the same distribution as first sampling a uniform ordered full-row-rank family R and then a uniform target index t. The predictor therefore inherits the learner's non-negligible advantage. The Goldreich–Levin hard-core theorem states precisely that a non-negligible predictor for $a^\star \cdot r^\star$ from $(F_n(a^\star), r^\star)$ yields an inverter for F_n with non-negligible probability. Thus such a learner would invert F_n (Goldreich and Levin, 1989), contradicting the assumed classical one-wayness. The training labels are thus not treated as advice: they are simulated inside the reduction itself. □

Proof of corollary 4. The first statement is exactly the constant-dimensional projected realization in corollary 5, which preserves the supervised concept class and uniform-distribution learning separation of Liu et al. (2021a). The second statement is corollary 6, obtained by applying the observable-to-interface compression in Lemma 7 to the hard observable family of Molteni et al. (2026). In both cases the cited work supplies the end-to-end quantum learner and classical learning lower bound; the interface constructions only change how the same target family is represented at the readout. □

Proof of corollary 5. We use the concept class of Liu et al. (2021a). Let p be prime, let g generate $\mathbb{Z}_p^*$, set $q = p - 1$, and write $\ell(x) = \log_g x \in \{0, \dots, q-1\}$ and $\theta_x = 2\pi\ell(x)/q$. A fault-tolerant quantum encoder can compute $\ell(x)$ in polynomial time using Shor's algorithm, amplified by majority voting to success probability $1 - \mathrm{negl}(n)$, apply $R_y(\theta_x)$ to one feature qubit, and uncompute the work register. The feature qubit is therefore

$$|\psi_x\rangle = \cos(\theta_x/2)|0\rangle + \sin(\theta_x/2)|1\rangle, \qquad \rho(x) = \frac{1}{2}\big(I + \sin\theta_x X + \cos\theta_x Z\big). \tag{40}$$

Only this qubit needs to reach the readout. It is thus a one-qubit valid causal separator in the sense of corollary 2; equivalently, it may be transferred to a blank readout by a SWAP, which uses three crossing CNOTs in the standard CNOT compilation. The populated Hermitian span is contained in $\mathrm{span}\{I, X, Z\}$ and, for $p \geq 5$, contains three linearly independent states, so $D = 3$. The corresponding linear projected kernel is

$$k(x, x') = \mathrm{Tr}[\rho(x)\rho(x')] = \frac{1}{2}\big(1 + \cos(\theta_x - \theta_{x'})\big). \tag{41}$$

The target concept f_s in Liu et al. (2021a, Eq. (2)) labels exactly the $q/2$ consecutive logarithms $\ell \in \{s, s+1, \dots, s+q/2-1\}$ modulo q by $+1$. Define

$$\phi_s = \frac{2\pi}{q}\left(s + \frac{q}{4} - \frac{1}{2}\right), \qquad W_s = \cos\phi_s\, Z + \sin\phi_s\, X. \tag{42}$$

The two zeros of $\cos(\theta - \phi_s)$ lie halfway between adjacent logarithm-grid points, hence

$$f_s(x) = \mathrm{sign}(\mathrm{Tr}[\rho(x)W_s]) = \mathrm{sign}(\cos(\theta_x - \phi_s))\,. \tag{43}$$

Thus every target concept is exactly a linear separator in the three-dimensional projected state span. After using Shor's algorithm to compute the angular coordinates of the labeled training examples, a quantum learner can fit a consistent half-circle separator in polynomial time and evaluate it on new examples by the same quantum feature computation. Standard VC/PAC bounds for two-dimensional linear separators give polynomial sample complexity. Classical hardness is exactly the hardness of the same concept class proved by Liu et al. (2021a, Theorem 1): under their discrete-log assumption, no probabilistic classical polynomial-time learner achieves accuracy $1/2 + 1/\mathrm{poly}(n)$ on the uniform distribution. The only new step here is the three-dimensional projected realization, so their lower bound transfers directly. □

Proof of Lemma 7. Prepare the quantum state $\rho_U(x)$ on the encoder and an index register J in the uniform state over $i \in [M]$. For each i, coherently implement the two-outcome measurement of P_i by a standard unitary dilation and store its eigenvalue sign in one outcome qubit. Because P_i has eigenvalues ± 1, the probability of observing sign $\sigma \in \{+1, -1\}$ conditioned on i is

$$\Pr[\sigma \mid i, x] = \frac{1 + \sigma\, \mu_i(x)}{2}, \qquad \mu_i(x) = \mathrm{Tr}[\rho_U(x) P_i]. \tag{44}$$

Consequently the joint branch label (i, σ) has probability

$$p_{i,\sigma}(x) = \frac{1 + \sigma\, \mu_i(x)}{2M}, \tag{45}$$

as claimed. Encode the at most $2M$ populated labels (i, σ) into $b_{\text{eff}} = \lceil \log_2(2M) \rceil$ control qubits, padding unused computational-basis labels with zero probability when necessary. A layer of control-to-readout CNOTs copies this branch code into an initially blank b_{eff}-qubit readout. After tracing out the encoder, the branch states on the readout are orthogonal computational-basis projectors, hence $G = I$ and $D \le 2M$. Finally,

$$p_{i,+}(x) - p_{i,-}(x) = \frac{\mu_i(x)}{M}, \tag{46}$$

so for any coefficient vector α,

$$f_\alpha(x) = \sum_{i=1}^{M} \alpha_i \mu_i(x) = M \sum_{i=1}^{M} \alpha_i \big(p_{i,+}(x) - p_{i,-}(x)\big). \tag{47}$$

Thus the observable-linear family lies exactly in the projected linear span. The claim concerns realizability and structural dimension; the norm of the representing readout functional can scale with $M\|\alpha\|$, so no margin-independent statistical bound is implied without additional assumptions on the coefficients and data distribution. □

Justification of corollary 6. Molteni et al. (2026, Theorem 15) construct a family of states $\rho_U(x)$ and an observable concept class $f_\alpha(x) = \mathrm{Tr}[\rho_U(x) O(\alpha)]$ with $O(\alpha) = \sum_{i=1}^{M} \alpha_i P_i$, where the number M of local Pauli terms is polynomial in the input size. Under their stated non-containment assumption for BQP and the relevant low level of the heuristic polynomial hierarchy, this family is identifiable by a quantum learner but not by the classical approximate-correct identification algorithms considered there. Applying Lemma 7 to the same $\rho_U(x)$ and $\{P_i\}$ produces a controlled projected representation with $b_{\text{eff}} = \lceil \log_2(2M) \rceil = O(\log n)$ and $D \le 2M = \mathrm{poly}(n)$, while preserving every target value $f_\alpha(x)$ exactly as a linear functional of the interface probabilities. This is a representation change, not a change of the underlying learning task, so the quantum identification algorithm and the classical hardness result of Molteni et al. (2026) transfer unchanged. As in Lemma 6, the interface is used to characterize the ideal representation and its width; the cited quantum identification algorithm is not asserted to operate through this projected readout, and the $O(1/M)$ branch probabilities make control-basis resolution of each μ_i cost $O(M^2 \varepsilon^{-2})$ samples. In particular, their separation arises from identifying the target concept and therefore goes beyond a lower bound based only on final hypothesis evaluation. □

E Finite-Rank Learning and Sample–Shot Proofs

This appendix collects the statistical learning results. It first establishes held-out interface-width selection, then gives rank-based and spectrum-aware kernel-ridge guarantees, and finally combines labeled-sample and interface-shot errors under the assumptions stated in the main text.

E.1 Width Selection

For a fit/validation split, write the generic held-out selector as

$$\hat{b} = \arg\min_{b \in \mathcal{B}} \left\{ \widehat{R}_{\text{val}}(\hat{h}_b) + \mathrm{pen}(b) \right\}. \tag{48}$$

Proposition 4 (Held-out width selection). *Fix Y with $|y| \le Y$ almost surely; clipping each $\hat{h}_b$ to $[-Y, Y]$ ensures that the squared losses lie in $[0, 4Y^2]$. For any nonnegative penalty with predictors trained independently of validation, with probability at least $1-\delta$,*

$$R(\hat{h}_{\hat{b}}) \le \min_{b \in \mathcal{B}} \{R(\hat{h}_b) + \mathrm{pen}(b)\} + 8Y^2 \sqrt{\frac{\log(2|\mathcal{B}|/\delta)}{2m_v}}. \tag{49}$$

Proof of Proposition 4. Condition on the training split, so the fitted predictors are fixed. Each squared loss lies in $[0, 4Y^2]$ by hypothesis. Hoeffding's inequality and a union bound over $\mathcal{B}$ give, simultaneously for all candidates and with probability at least $1-\delta$,

$$|\widehat{R}_{\mathrm{val}}(\hat{h}_b) - R(\hat{h}_b)| \le 4Y^2 \sqrt{\frac{\log(2|\mathcal{B}|/\delta)}{2m_v}} =: \epsilon_v. \tag{50}$$

For any $b \in \mathcal{B}$, the optimality of Eq. (48) and $\mathrm{pen}(\hat{b}) \ge 0$ then imply

$$R(\hat{h}_{\hat{b}}) \le \widehat{R}_{\mathrm{val}}(\hat{h}_{\hat{b}}) + \epsilon_v \tag{51}$$

$$\le \widehat{R}_{\mathrm{val}}(\hat{h}_b) + \mathrm{pen}(b) - \mathrm{pen}(\hat{b}) + \epsilon_v \tag{52}$$

$$\le R(\hat{h}_b) + \mathrm{pen}(b) + 2\epsilon_v. \tag{53}$$

Taking the minimum over b yields Eq. (49). The argument uses only the nonnegativity and split-independence of the penalty. It therefore validates selection for any such penalty; it does not show that the particular choice proportional to $2^b/m_f$ improves on validation alone. □

Extension to Gaussian and conditionally sub-Gaussian labels. The bounded-response condition in Proposition 4 can be replaced by $y = f^\circ(x) + \xi$, $|f^\circ(x)| \le F$, and $\mathbb{E}[e^{t\xi} \mid x] \le e^{\sigma^2 t^2/2}$ for every real t. Conditionally on the fit split, suppose all candidates satisfy $|\hat{h}_b(x)| \le H$ and the nonnegative penalty is fixed without validation data. Then Eq. (49) holds with its final term replaced by

$$2\sqrt{\frac{2v\log(2|\mathcal{B}|/\delta)}{m_v}}, \qquad v = \frac{(H^2+4HF)^2}{4} + 4H^2\sigma^2. \tag{54}$$

This extension includes independent Gaussian label noise without requiring bounded observed labels.

To prove the claim, subtract the same observed baseline loss y^2 from every candidate's squared loss. This subtraction leaves the selected width unchanged and gives $d_h = h(x)^2 - 2h(x)f^\circ(x) - 2h(x)\xi$. The first two terms lie in $[-2HF, H^2+2HF]$. Conditioned first on x, the noise term has a moment-generating function of at most $\exp(2t^2H^2\sigma^2)$. Hoeffding's lemma for the remaining bounded term therefore yields $\mathbb{E}\exp[t(d_h - \mathbb{E}d_h)] \le \exp(t^2 v/2)$. A union bound gives $|\widehat{d}_h - \mathbb{E}d_h| \le \sqrt{2v\log(2|\mathcal{B}|/\delta)/m_v}$ for every candidate. Repeating the preceding selection argument for these centered losses proves Eq. (54), since $R(0)$ cancels from both sides. The result controls the selected fit-split predictor; a subsequent refit is a different estimator, compared separately in Section H.3.4. If the penalties themselves are validation-dependent, a valid alternative is to select from a finite, prespecified family of candidate–penalty pairs: the same argument uses that family's cardinality in place of $|\mathcal{B}|$. For a full product with identical predictors and penalties $\lambda 2^b/m_f$, this joint minimization simply chooses the smallest allowed λ; it does not justify an arbitrary validation-derived scale.

Closed-form oracle under exponential approximation. For $A_0, C, \nu > 0$ and $m > 0$, consider the approximation–estimation surrogate $A_0 e^{-\nu b} + C2^b/m$. Differentiating this strictly convex function of b gives the continuous minimizer

$$b^*_{\mathrm{cont}} = \frac{\log\big(\nu A_0 m/(C\log 2)\big)}{\nu + \log 2}. \tag{55}$$

If b^*_{cont} belongs to the candidate grid, it also minimizes the surrogate on that grid. Otherwise, compare the closest feasible grid points below and above it, using the corresponding endpoint when

the continuous minimizer lies outside the grid. This gives a logarithmic optimizer of the surrogate; when the approximation envelope is unavailable, Proposition 4 provides a data-driven selector with an oracle guarantee.

E.2 Finite-Rank KRR Conditions and Excess Risk

For vectors u, v in a Hilbert space, write $u \boxtimes v : T \mapsto \langle v, T\rangle u$ for their rank-one operator. This ordering makes $u \boxtimes v$ linear under the convention that the inner product is linear in its second argument.

Assumption 1 (Finite-rank KRR conditions)**.** *Fix a width b.*

The response is real with a finite second moment, the loss is squared error, and the width-b RKHS has a minimum-norm population risk minimizer $h_b^\star$. Write $f^\circ(x) = \mathbb{E}[y \mid x]$ and $A_b = R(h_b^\star) - R(f^\circ)$. Let $\varphi_b(x) = \rho_B(x)$, $\Sigma_b = \mathbb{E}[\varphi_b(x) \boxtimes \varphi_b(x)]$, $D_b = \operatorname{rank}(\Sigma_b)$, and $\mathcal{V}_b = \operatorname{range}(\Sigma_b)$, viewed as a D_b-dimensional real Hilbert space of Hermitian operators. Let $w_b^\star \in \mathcal{V}_b$ be the minimum-norm coefficient of $h_b^\star$, and use Σ_b^{-1} only on $\mathcal{V}_b$. Assume $x \mapsto \varphi_b(x)$ is continuous on $\operatorname{supp}(P)$, as required by Lemma 1. Also assume that $\Lambda_b = \sup_x \|\varphi_b(x)\|_{\mathrm{HS}}^2 \le 1$ and that the smallest nonzero eigenvalue of Σ_b is at least $\mu_b > 0$.

For the residual $\zeta_b = y - h_b^\star(x)$, assume the whitened feature-weighted score

$$g_b(x, y) = \zeta_b \, \Sigma_b^{-1/2} \varphi_b(x) \tag{56}$$

is centered and sub-Gaussian with proxy σ_b^2, meaning that, for every $u \in \mathcal{V}_b$,

$$\mathbb{E} \exp\big(\langle u, g_b\rangle_{\mathbb{R}}\big) \le \exp\big(\sigma_b^2 \|u\|^2/2\big). \tag{57}$$

Here $\langle X, Y\rangle_{\mathbb{R}} = \operatorname{Tr}[XY]$ denotes the trace inner product on Hermitian operators. The centering $\mathbb{E}[\zeta_b \varphi_b(x)] = 0$ is the normal equation for $h_b^\star$; the tail inequality is an additional assumption, not a consequence of finite rank.

We further require the sample to resolve the populated feature space:

$$m \;\ge\; c \, \frac{\Lambda_b}{\mu_b} \, \log \frac{2 D_b}{\delta} \tag{58}$$

for a universal constant c; $c = 20$ is sufficient in the proof below.

Alternatively, replace the score condition by the well-specified model $f^\circ = h_b^\star$ and conditionally centered scalar noise $\xi = y - f^\circ(x)$ satisfying $\mathbb{E}[e^{t\xi} \mid x] \le e^{\sigma^2 t^2/2}$ for every real t. This alternative retains all feature and sample-resolution conditions but needs no unconditional sub-Gaussian bound on the whitened score.

The inverse and the probability inequalities in Assumption 1 are thus applied in D_b real Hermitian coordinates. In these coordinates, the covariance is represented by a real symmetric positive-semidefinite matrix.

Remark 2 (The covariance floor is forced to shrink). Since $\varphi_b(x)$ is a density matrix, $\operatorname{Tr} \Sigma_b = \mathbb{E}\|\varphi_b(x)\|_{\mathrm{HS}}^2 = \mathbb{E} \operatorname{Tr}[\rho_B(x)^2] \le \Lambda_b \le 1$. By Lemma 1, the operator Σ_b has exactly D_b nonzero eigenvalues, each at least μ_b, so $D_b \mu_b \le \operatorname{Tr} \Sigma_b \le \Lambda_b$ and therefore

$$\mu_b \;\le\; \frac{\Lambda_b}{D_b}. \tag{59}$$

The covariance floor is therefore not merely encoder dependent: it must shrink at least as fast as $1/D_b$. Equality requires the D_b populated directions to carry equal weight, which rank saturation does not imply, since rank counts populated directions and says nothing about their weights. Substituting Eq. (59) into Eq. (58) shows that Assumption 1 is never satisfied below

$$m \;\gtrsim\; D_b \log(2 D_b/\delta). \tag{60}$$

Equation (60) is necessary but not sufficient: Eq. (59) bounds μ_b from above, and μ_b may lie far below this upper bound, so certifying Eq. (58) from D_b alone would require a lower bound on μ_b,

which is not available. Under the two further hypotheses that the controlled ceiling is saturated, $D_b = \Theta(2^b)$, and that the populated spectrum is balanced, $\mu_b = \Theta(\Lambda_b/D_b)$, the condition scales as

$$m \gtrsim 2^b\big(b + \log(2/\delta)\big), \quad \text{i.e.} \quad b \le \log_2 m - \log_2\big(b + \log(2/\delta)\big) - O(1), \tag{61}$$

This scaling gives a log-logarithmic rather than additive correction to the width accounting of Section 3.2. The balanced construction in Section H.2.3 has $D_b = 2^b$ and $\mu_b = 1/D_b$ exactly and uses sample sizes satisfying Eq. (58) with $c = 20$. In contrast, the image and XOR width sweeps do not establish population spectral balance or the covariance-resolution condition. Their finite-Gram rank diagnostics count directions but do not measure the population covariance floor.

Proposition 5 (Finite-rank KRR excess risk)**.** *Let Assumption 1 hold and let $\gamma > 0$. Then, with probability at least $1 - \delta$,*

$$R(\hat{h}_{\gamma,b}) - R(f^\circ) \le A_b + 4\gamma\|h_b^\star\|^2_{\mathcal{H}_{k_B}} + C_2\frac{D_b + \log(2/\delta)}{m}. \tag{62}$$

Under the score condition, C_2 depends only on σ_b^2; under the well-specified conditional-noise alternative, one may take $C_2 = 36\sigma^2$. The covariance floor enters through Eq. (58), with any dependence carried by the relevant noise or score proxy made explicit. Under Eq. (3), corollary 1 gives $D_b \le 2^b$, so the estimation contribution is of order $2^b/m$ when those proxies are uniformly bounded across the compared widths.

Proof. Abbreviate $\varphi_b, \Sigma_b, D_b, \mu_b, \Lambda_b, w_b^\star$ by $\varphi, \Sigma, D, \mu, \Lambda, w^\star$, and work in the D-dimensional real Hermitian space $\mathcal{V} = \mathrm{range}(\Sigma)$. In these coordinates the minimum-norm coefficient satisfies $h_b^\star(x) = \langle\varphi(x), w^\star\rangle_{\mathbb{R}}$. Put

$$\widehat{\Sigma} = \frac{1}{m}\sum_{i=1}^m \varphi(x_i) \boxtimes \varphi(x_i), \qquad v = \frac{1}{m}\sum_{i=1}^m \zeta_{b,i}\varphi(x_i), \qquad M = \widehat{\Sigma} + \gamma I_{\mathcal{V}}. \tag{63}$$

The primal form of KRR and $y_i = h_b^\star(x_i) + \zeta_{b,i}$ give the exact identity

$$\widehat{w}_{\gamma,b} - w^\star = M^{-1}(v - \gamma w^\star). \tag{64}$$

We first control the empirical covariance relative to its population counterpart. Put $Z_i = \Sigma^{-1/2}\varphi(x_i)$. Then $\mathbb{E}[Z \boxtimes Z] = I_{\mathcal{V}}$ and $\|Z_i\|^2 \le \Lambda/\mu$ almost surely. The matrix Chernoff inequality (Tropp, 2015, Thm. 5.1.1), applied in these real coordinates, gives

$$\Pr\Big[\big\|m^{-1}\textstyle\sum_i Z_i \boxtimes Z_i - I_{\mathcal{V}}\big\| > \frac{1}{2}\Big] \le 2D\exp\Big(-\tfrac{m\mu}{10\Lambda}\Big) \le \tfrac{\delta}{2}, \tag{65}$$

for a sufficiently large universal constant c in Eq. (58). Call the complement of the event inside the probability $\mathcal{E}_1$. On $\mathcal{E}_1$, $\frac{1}{2}\Sigma \preceq \widehat{\Sigma} \preceq \frac{3}{2}\Sigma$.

We next control the stochastic score. By Eq. (57), the independent mean-zero vectors $g_{b,i} = \zeta_{b,i}Z_i$ have sub-Gaussian proxy σ_b^2. Their average is $\Sigma^{-1/2}v$. The sub-Gaussian quadratic-form inequality of Hsu et al. (2012), applied unconditionally with the identity matrix in the D real coordinates, yields an event $\mathcal{E}_2$ of probability at least $1 - \delta/2$ on which

$$\|\Sigma^{-1/2}v\|^2 \le C\sigma_b^2\frac{D + \log(2/\delta)}{m}, \tag{66}$$

where C is an absolute constant.

For the conditional-noise alternative, fix the inputs and form the matrix

$$A = m^{-1}[Z_1|\cdots|Z_m].$$

On $\mathcal{E}_1$, $AA^\top \preceq 3I/(2m)$, while the conditionally independent noise variables have joint sub-Gaussian proxy σ^2. The same quadratic-form inequality, now applied conditionally to $A\xi = \Sigma^{-1/2}v$, gives for $t = \log(2/\delta)$

$$\|\Sigma^{-1/2}v\|^2 \le \frac{3\sigma^2}{2m}(D + 2\sqrt{Dt} + 2t) \le \frac{9\sigma^2}{2m}(D + t). \tag{67}$$

The failure probability of this bound on $\mathcal{E}_1$ is at most $\delta/2$, so the joint covariance-and-noise event still has probability at least $1-\delta$.

We now translate these two concentration events into excess risk. The excess risk is measured in the covariance norm

$$\|u\|_\Sigma^2 := \langle u, \Sigma u\rangle_{\mathbb{R}}, \qquad u \in \mathcal{V}.$$

On $\mathcal{E}_1$, put $B = \Sigma^{-1/2} M \Sigma^{-1/2} \succeq \frac{1}{2} I_{\mathcal{V}}$. Therefore

$$\|M^{-1}v\|_\Sigma = \|B^{-1}\Sigma^{-1/2}v\| \le 2\|\Sigma^{-1/2}v\|. \tag{68}$$

Moreover, $\Sigma \preceq 2M$ and $M \succeq \gamma I_{\mathcal{V}}$ imply, by congruence rather than by commuting the two covariance matrices,

$$\gamma^2\|M^{-1}w^\star\|_\Sigma^2 \le 2\gamma^2\langle w^\star, M^{-1}w^\star\rangle_{\mathbb{R}} \le 2\gamma\|w^\star\|^2. \tag{69}$$

Combining these inequalities with Eq. (64) gives, on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$R(\widehat{h}_{\gamma,b}) - R(h_b^\star) = \|\widehat{w}_{\gamma,b} - w^\star\|_\Sigma^2 \le 4\gamma\|h_b^\star\|_{\mathcal{H}_{k_B}}^2 + 8C\sigma_b^2\frac{D_b + \log(2/\delta)}{m}. \tag{70}$$

The equality is the population normal equation for the best-in-RKHS predictor; it does not require f° to belong to the RKHS. Finally, $R(h_b^\star) - R(f^\circ) = A_b$, so adding the approximation term proves Eq. (62) with $C_2 = 8C\sigma_b^2$. The union bound gives probability at least $1-\delta$. Only the event $\mathcal{E}_1$ depends on μ_b, through the sample-resolution condition, and substituting $D_b \le 2^b$ proves the last statement. Using Eq. (67) in the same excess-risk argument instead gives $C_2 = 36\sigma^2$ and $A_b = 0$ under the well-specified alternative. These two routes are distinct: scalar Gaussian noise alone does not imply a width-independent unconditional score proxy in a misspecified model. □

E.3 Spectrum-Aware KRR Excess Risk

Proposition 6 (Spectrum-aware KRR excess risk). *Fix a width b and retain the setting and feature conditions of Assumption 1. Assume either the whitened-score condition Eq. (57), or the well-specified alternative $f^\circ = h_b^\star$ with conditionally centered scalar noise $\xi = y - f^\circ(x)$ satisfying $\mathbb{E}[e^{t\xi} \mid x] \le e^{\sigma^2 t^2/2}$. Replace the covariance-floor sample condition by the regularized leverage condition below. For $\gamma > 0$, define*

$$\ell_{\gamma,b} := \sup_x \left\langle \varphi_b(x), (\Sigma_b + \gamma I)^{-1}\varphi_b(x)\right\rangle_{\mathbb{R}}, \qquad d_{\mathrm{eff},b}(\gamma) := \mathrm{Tr}\left[\Sigma_b(\Sigma_b + \gamma I)^{-1}\right]. \tag{71}$$

Assume the regularized geometry is resolved in the sense that

$$m \ge c\,(1 + \ell_{\gamma,b})\log\frac{4D_b}{\delta} \tag{72}$$

for a sufficiently large universal constant c. Under the whitened-score condition, with probability at least $1-\delta$,

$$R(\hat{h}_{\gamma,b}) - R(f^\circ) \le A_b + 8\gamma\|h_b^\star\|_{\mathcal{H}_{k_B}}^2 + 24\sigma_b^2\frac{d_{\mathrm{eff},b}(\gamma) + \log(4/\delta)}{m}. \tag{73}$$

Under the well-specified conditional-noise alternative, the same effective-dimension dependence holds without an unconditional whitened-score proxy:

$$R(\hat{h}_{\gamma,b}) - R(f^\circ) \le 8\gamma\|h_b^\star\|_{\mathcal{H}_{k_B}}^2 + 72\sigma^2\frac{d_{\mathrm{eff},b}(\gamma) + \log(4/\delta)}{m}. \tag{74}$$

For the controlled interface, corollary 3 identifies $d_{\mathrm{eff},b}(\gamma)$ with the regularized effective dimension of the finite matrix $M_b = G_b^{1/2}C_{p,b}G_b^{1/2}$. Since $d_{\mathrm{eff},b}(\gamma) \le D_b \le 2^b$, Eq. (73) refines the rank-based envelope whenever the populated spectrum is nonuniform at the learning scale γ.

Proof. Abbreviate $\varphi_b, \Sigma_b, D_b, w_b^\star$ by $\varphi, \Sigma, D, w^\star$ and work on $\mathcal{V} = \mathrm{range}(\Sigma)$. We first resolve the covariance at the regularization scale. Let

$$\Gamma = \Sigma + \gamma I_{\mathcal{V}}, \qquad \widehat{\Gamma} = \widehat{\Sigma} + \gamma I_{\mathcal{V}}, \qquad A = \Gamma^{-1/2}\Sigma\Gamma^{-1/2}.$$

Then $0 \preceq A \preceq I$ and $\operatorname{Tr} A = d_{\mathrm{eff},b}(\gamma)$. For $Z_i = \Gamma^{-1/2}\varphi(x_i)$, Eq. (71) gives $\|Z_i\|^2 \le \ell_{\gamma,b}$ and $\mathbb{E}[Z \boxtimes Z] = A$. The self-adjoint matrix Bernstein inequality (Tropp, 2015), applied to $Z_i \boxtimes Z_i - A$, implies that under Eq. (72), with probability at least $1 - \delta/2$,

$$\left\|\Gamma^{-1/2}(\widehat{\Sigma} - \Sigma)\Gamma^{-1/2}\right\| \le \frac{1}{2}. \tag{75}$$

On this event,

$$B := \Gamma^{-1/2}\widehat{\Gamma}\Gamma^{-1/2} = I + \Gamma^{-1/2}(\widehat{\Sigma} - \Sigma)\Gamma^{-1/2} \succeq \tfrac{1}{2}I.$$

Under the whitened-score assumption, we next control the stochastic term. Recall $v = m^{-1}\sum_i \zeta_{b,i}\varphi(x_i)$ from the proof of Proposition 5. Under the whitened-score condition, $g_i = \zeta_{b,i}\Sigma^{-1/2}\varphi(x_i)$ is centered sub-Gaussian with proxy $\sigma_b^2 I$ on $\mathcal{V}$. Writing

$$\Gamma^{-1/2}v = \frac{1}{m}\sum_{i=1}^{m} T g_i, \qquad T = \Gamma^{-1/2}\Sigma^{1/2},$$

shows that this average is sub-Gaussian with covariance proxy $\sigma_b^2 A/m$. The sub-Gaussian quadratic-form inequality of Hsu et al. (2012) therefore yields, with probability at least $1 - \delta/2$ and $t = \log(4/\delta)$,

$$\|\Gamma^{-1/2}v\|^2 \le \frac{\sigma_b^2}{m}\left(\operatorname{Tr} A + 2\sqrt{\operatorname{Tr}(A^2)t} + 2\|A\|t\right) \tag{76}$$

$$\le 3\sigma_b^2 \frac{d_{\mathrm{eff},b}(\gamma) + t}{m}, \tag{77}$$

where $0 \preceq A \preceq I$ implies $\operatorname{Tr}(A^2) \le \operatorname{Tr} A$ and $\|A\| \le 1$.

For the conditional-noise alternative, the same role is played by a conditional quadratic-form bound. Put $\widehat{A} = m^{-1}\sum_i Z_i \boxtimes Z_i$ and $\widehat{d} = \operatorname{Tr}\widehat{A} = m^{-1}\sum_i \|Z_i\|^2$. The variables $\|Z_i\|^2$ lie in $[0, \ell_{\gamma,b}]$, have mean $d_{\mathrm{eff},b}(\gamma)$, and variance at most $\ell_{\gamma,b} d_{\mathrm{eff},b}(\gamma)$. Scalar Bernstein concentration and Eq. (72) therefore give, with probability at least $1 - \delta/4$,

$$\widehat{d} \le d_{\mathrm{eff},b}(\gamma) + \sqrt{\frac{2\ell_{\gamma,b} d_{\mathrm{eff},b}(\gamma)t}{m}} + \frac{\ell_{\gamma,b}t}{3m} \le 2\big(d_{\mathrm{eff},b}(\gamma) + t\big), \tag{78}$$

for a sufficiently large universal c. Condition on the sampled inputs. With the matrix $C = m^{-1}[Z_1|\cdots|Z_m]$, conditional independence and sub-Gaussianity of the scalar noises give $\Gamma^{-1/2}v = C\xi$ and $CC^\top = \widehat{A}/m$. On Eq. (75), $\|\widehat{A}\| \le 3/2$ and hence $\operatorname{Tr}(\widehat{A}^2) \le (3/2)\widehat{d}$. Applying the same quadratic-form inequality conditionally and using Eq. (78), with conditional failure probability at most $\delta/4$,

$$\|\Gamma^{-1/2}v\|^2 \le \frac{\sigma^2}{m}\left(\widehat{d} + 2\sqrt{\operatorname{Tr}(\widehat{A}^2)t} + 2\|\widehat{A}\|t\right) \tag{79}$$

$$\le 9\sigma^2 \frac{d_{\mathrm{eff},b}(\gamma) + t}{m}. \tag{80}$$

Thus conditional scalar noise produces the same effective-dimension scaling; no covariance-floor or unconditional whitened-score bound is needed.

It remains to convert the score bounds into excess risk. The primal identity Eq. (64) remains $\widehat{w}_{\gamma,b} - w^\star = \widehat{\Gamma}^{-1}(v - \gamma w^\star)$. Because $\Sigma \preceq \Gamma$ and $B \succeq I/2$,

$$\|\widehat{\Gamma}^{-1}v\|_\Sigma^2 \le \|\widehat{\Gamma}^{-1}v\|_\Gamma^2 = \|B^{-1}\Gamma^{-1/2}v\|^2 \le 4\|\Gamma^{-1/2}v\|^2, \tag{81}$$

$$\gamma^2\|\widehat{\Gamma}^{-1}w^\star\|_\Sigma^2 \le 4\gamma^2\langle w^\star, \Gamma^{-1}w^\star\rangle_{\mathbb{R}} \le 4\gamma\|w^\star\|^2. \tag{82}$$

Using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and Eq. (77) gives

$$R(\widehat{h}_{\gamma,b}) - R(h_b^\star) \le 8\gamma\|h_b^\star\|_{\mathcal{H}_{k_B}}^2 + 24\sigma_b^2\frac{d_{\mathrm{eff},b}(\gamma) + t}{m},$$

which proves Eq. (73) after adding A_b. Under the well-specified alternative, use Eq. (80) instead; the same deterministic resolvent step multiplies its stochastic term by at most eight, giving Eq. (74) and $A_b = 0$. For the conditional-noise route, allocate failure probability $\delta/2$ to the covariance event and $\delta/4$ each to the empirical-trace and conditional quadratic-form events. For the whitened-score route, allocate $\delta/2$ to each of the covariance and score events. Thus the relevant events have total failure probability at most δ; the slightly stronger logarithm in Eq. (72) absorbs this bookkeeping. Finally, corollary 3 equates the nonzero spectrum of Σ_b (equivalently L_{k_B}) with that of M_b in the controlled representation, giving the stated finite-matrix form. □

E.4 Joint Labeled-Sample and Interface-Shot Resources

Proposition 7 (Joint labeled-sample and interface-shot excess risk)**.** *Assume the controlled interface of Eq. (3), known branch geometry G, and the conditions of Proposition 6 with confidence parameter $\delta/2$. Assume in addition that $|y| \le Y$ almost surely. For $g_s = G^{1/2} e_s$, let*

$$\hat{z}_S(x) = \frac{1}{S} \sum_{t=1}^{S} g_{s_t}, \qquad s_t \overset{\text{iid}}{\sim} p(x), \tag{83}$$

and fit ridge regression with regularization γ on the m acquired training features $\hat{z}_S(x_i)$. At prediction time, use an independent S-sample estimate Eq. (83) for the query. Denote the resulting randomized predictor by $\tilde{h}_{\gamma,b,S}$ and its population risk averaged over fresh query acquisition by

$$\overline{R}_S(\tilde{h}_{\gamma,b,S}) := \mathbb{E}_{x,y} \mathbb{E}_{\text{query shots}} \big[(\tilde{h}_{\gamma,b,S}(x) - y)^2 \big]. \tag{84}$$

Define

$$B_{m,b}(\gamma, \delta) := A_b + 8\gamma \|h_b^\star\|^2_{\mathcal{H}_{k_B}} + 24\sigma_b^2 \frac{d_{\text{eff},b}(\gamma) + \log(8/\delta)}{m}, \tag{85}$$

$$\varepsilon_{S,m,\delta} := \frac{1 + \sqrt{2\log(4m/\delta)}}{\sqrt{S}}, \tag{86}$$

$$\Delta_{S,m}(\gamma, \delta)^2 := \frac{Y^2}{\gamma^2} \left[\left(1 + \frac{2}{\gamma} \right)^2 \varepsilon^2_{S,m,\delta} + \frac{1}{S} \right]. \tag{87}$$

Then, with probability at least $1 - \delta$ over the labeled training sample and the training acquisition shots,

$$\sqrt{\overline{R}_S(\tilde{h}_{\gamma,b,S}) - R(f^\circ)} \le \sqrt{B_{m,b}(\gamma, \delta)} + \Delta_{S,m}(\gamma, \delta). \tag{88}$$

Consequently, for $\gamma \le 1$,

$$\overline{R}_S(\tilde{h}_{\gamma,b,S}) - R(f^\circ) \le 2B_{m,b}(\gamma, \delta) + C \frac{Y^2 \big(1 + \log(4m/\delta)\big)}{\gamma^4 S} \tag{89}$$

for a universal constant C.

Proof. Because G is the Hilbert–Schmidt Gram matrix of the branch states, $G \succeq 0$ and

$$\|g_s\|^2 = e_s^\top G e_s = G_{ss} = \operatorname{Tr}(\tau_s^2) \le 1.$$

Moreover, $z(x) = G^{1/2} p(x) = \mathbb{E}[g_s \mid x]$ and $\|z(x)\| \le 1$. Thus Eq. (83) is an unbiased finite-dimensional feature estimator with $\|\hat{z}_S(x)\| \le 1$. Its mean-square error obeys

$$\mathbb{E}\|\hat{z}_S(x) - z(x)\|^2 = \frac{1}{S} \mathbb{E}\|g_s - z(x)\|^2 \le \frac{1}{S}. \tag{90}$$

Also, $\mathbb{E}\|\hat{z}_S - z\| \le S^{-1/2}$ by Jensen. Changing one of the S samples changes the norm $\|\hat{z}_S - z\|$ by at most $2/S$. McDiarmid's inequality therefore gives

$$\Pr\left[\|\hat{z}_S(x) - z(x)\| > \frac{1 + \sqrt{2\log(1/\eta)}}{\sqrt{S}} \right] \le \eta.$$

Applying this bound with $\eta = \delta/(2m)$ and a union bound over the m training inputs yields an event $\mathcal{E}_S$ of probability at least $1 - \delta/2$ on which

$$\max_{i\in[m]} \|\hat{z}_S(x_i) - z(x_i)\| \le \varepsilon_{S,m,\delta}. \tag{91}$$

Work now in the explicit interface feature space with exact empirical covariance and response vector

$$A = \frac{1}{m}\sum_{i=1}^{m} z_i z_i^\top + \gamma I, \qquad b = \frac{1}{m}\sum_{i=1}^{m} y_i z_i, \qquad w = A^{-1}b,$$

and their shot-acquired counterparts $\tilde{A}, \tilde{b}, \tilde{w}$ obtained by replacing z_i by $\hat{z}_i$. The exact predictor $h(x) = w^\top z(x)$ is the same KRR estimator as $\hat{h}_{\gamma,b}$ because $k_B(x, x') = z(x)^\top z(x')$. Since $A, \tilde{A} \succeq \gamma I$, $\|z_i\|, \|\hat{z}_i\| \le 1$, and $|y_i| \le Y$,

$$\|w\| \le \frac{Y}{\gamma}, \qquad \|\tilde{w}\| \le \frac{Y}{\gamma}. \tag{92}$$

On $\mathcal{E}_S$,

$$\|\tilde{b} - b\| \le Y\varepsilon_{S,m,\delta}, \tag{93}$$

$$\|\tilde{A} - A\| \le \frac{1}{m}\sum_i \|\hat{z}_i\hat{z}_i^\top - z_i z_i^\top\| \le 2\varepsilon_{S,m,\delta}. \tag{94}$$

Using $Aw = b$ and $\tilde{A}\tilde{w} = \tilde{b}$ therefore gives

$$\|\tilde{w} - w\| \le \|\tilde{A}^{-1}\| \left(\|\tilde{b} - b\| + \|\tilde{A} - A\|\,\|w\|\right) \tag{95}$$

$$\le \frac{Y\varepsilon_{S,m,\delta}}{\gamma}\left(1 + \frac{2}{\gamma}\right). \tag{96}$$

For a fresh query x, let $\hat{z} = \hat{z}_S(x)$ be acquired independently of the training shots. Conditional on the trained coefficients, $\mathbb{E}[\hat{z} - z(x) \mid x] = 0$. Hence the cross term vanishes and, using Eqs. (90), (92) and (96),

$$\mathbb{E}_{\text{query shots}}\left[(\tilde{w}^\top\hat{z} - w^\top z(x))^2\right] = \left((\tilde{w} - w)^\top z(x)\right)^2 + \mathbb{E}\left[(\tilde{w}^\top(\hat{z} - z(x)))^2\right] \tag{97}$$

$$\le \frac{Y^2\varepsilon_{S,m,\delta}^2}{\gamma^2}\left(1 + \frac{2}{\gamma}\right)^2 + \frac{Y^2}{\gamma^2 S} \tag{98}$$

$$= \Delta_{S,m}(\gamma,\delta)^2. \tag{99}$$

The same bound holds after averaging over $x \sim P$.

Apply Proposition 6 with confidence parameter $\delta/2$. Its event $\mathcal{E}_L$ has probability at least $1 - \delta/2$ and gives $R(h) - R(f^\circ) \le B_{m,b}(\gamma,\delta)$. On $\mathcal{E}_L \cap \mathcal{E}_S$, which has probability at least $1 - \delta$, Minkowski's inequality in $L^2(P \times \text{query shots})$ and Eq. (99) give

$$\sqrt{\overline{R}_S(\tilde{h}_{\gamma,b,S}) - R(f^\circ)} = \|\tilde{h}_{\gamma,b,S} - f^\circ\|_{L^2(P\times\text{shots})} \tag{100}$$

$$\le \|h - f^\circ\|_{L^2(P)} + \|\tilde{h}_{\gamma,b,S} - h\|_{L^2(P\times\text{shots})} \tag{101}$$

$$\le \sqrt{B_{m,b}(\gamma,\delta)} + \Delta_{S,m}(\gamma,\delta), \tag{102}$$

proving Eq. (88). Finally, $(a+b)^2 \le 2a^2 + 2b^2$, $\gamma \le 1$, and $(1 + \sqrt{2u})^2 \le 2 + 4u$ for $u \ge 0$ give Eq. (89) after absorbing numerical constants into C. □

Corollary 7 (Training-budget balance)**.** *Fix b and $\gamma \le 1$, and let $\mathcal{B} = mS$ be the number of controlled-interface state preparations used to acquire the m training features, ignoring the external*

cost of labels. Suppressing confidence logarithms and fixed approximation/regularization terms, Eq. (89) has stochastic part

$$\widetilde{O}\Big(\frac{a_b}{m}+\frac{c_\gamma}{S}\Big), \qquad a_b=\sigma_b^2 d_{\mathrm{eff},b}(\gamma), \quad c_\gamma=\frac{Y^2}{\gamma^4}. \tag{103}$$

Under $mS=\mathcal{B}$, and provided the balance point is feasible, $1\le m_\star\le\mathcal{B}$, the two terms are balanced at

$$m_\star=\widetilde{\Theta}\left(\sqrt{\frac{a_b\mathcal{B}}{c_\gamma}}\right), \qquad S_\star=\widetilde{\Theta}\left(\sqrt{\frac{c_\gamma\mathcal{B}}{a_b}}\right), \tag{104}$$

with stochastic excess

$$\widetilde{O}\left(\sqrt{\frac{a_b c_\gamma}{\mathcal{B}}}\right)=\widetilde{O}\left(\frac{\sigma_b Y}{\gamma^2}\sqrt{\frac{d_{\mathrm{eff},b}(\gamma)}{\mathcal{B}}}\right). \tag{105}$$

Proof. Substitute $S=\mathcal{B}/m$ into Eq. (103) to obtain, up to logarithmic factors, $a_b/m+c_\gamma m/\mathcal{B}$. The two positive terms are minimized at $m=\sqrt{a_b\mathcal{B}/c_\gamma}$, with $S=\mathcal{B}/m$; substituting this choice yields Eq. (105). □

F Finite-Measurement and Sharp Interface Results

This appendix collects the finite-measurement and sharp-interface results used by the main text and the experiments. The statements and their proofs are kept together here; Appendix H contains only their numerical instantiations and empirical checks.

F.1 Structural Capacity under Local Interleaving

The controlled-sector form in Eq. (3) yields the 2^b ceiling of corollary 1. For binary A-controlled crossings with arbitrary fixed local interleaving, the following prefix-span argument improves the generic 4^b bound. Assume every crossing is binary controlled from A, $C_j=P_{j,0}\otimes I+P_{j,1}\otimes U_j$, with complementary orthogonal encoder projectors; the projectors need not agree between crossings. Between successive crossings, arbitrary fixed local unitaries may act independently on either register.

Proposition 8 (Prefix-span bound under local interleaving)**.** *For a pure, data-independent initial readout state and b binary A-controlled crossings with arbitrary fixed interleaved local unitaries on either register,*

$$D_b\le 1+\sum_{j=1}^{b}4^{j-1}=\frac{4^b+2}{3}. \tag{106}$$

This bound applies to CNOT interfaces with control-mixing local gates, without requiring a common orthogonal-sector decomposition.

Proof of Proposition 8. Let D_j denote the span dimension after the first j crossings and adjacent local operations, with $D_0=1$. Before crossing j, the operator-Schmidt expansion of the prefix places every joint state in $\mathcal{H}_A\otimes\mathcal{S}_{j-1}$ for a fixed readout subspace $\mathcal{S}_{j-1}\subseteq\mathcal{H}_B$ of Hilbert-space dimension at most 2^{j-1}. Local encoder unitaries leave the reduced state unchanged, and local readout unitaries rotate this subspace without changing its dimension. For the joint state $\omega(x)$ immediately before crossing j, put $\sigma_a(x)=\mathrm{Tr}_A[(P_{j,a}\otimes I)\omega(x)(P_{j,a}\otimes I)]$ for $a\in\{0,1\}$. These are subnormalized operators on $\mathcal{S}_{j-1}$, and $\sigma_0+\sigma_1=\rho_{j-1}$ because the partial trace cancels the off-diagonal projector blocks. The new readout is therefore

$$\rho_j(x)=\sigma_0(x)+U_j\sigma_1(x)U_j^\dagger=\rho_{j-1}(x)+(\mathrm{Ad}_{U_j}-I)\sigma_1(x). \tag{107}$$

The first term spans D_{j-1} dimensions and the second is the image of a subspace of $\mathrm{HS}(\mathcal{S}_{j-1})$, which has operator dimension at most $(2^{j-1})^2=4^{j-1}$. Thus $D_j\le D_{j-1}+4^{j-1}$; summing from $D_0=1$ proves the result. The conditional operators may retain coherences, so the argument bounds their operator-span contribution without assuming a classical-interface factorization. □

F.2 Exact and Minimax Learning Costs

The next two results make the finite-sample cost of a saturated controlled interface explicit. They complement the general KRR bounds in Section E.1 by treating a categorical copy interface for which the population risk can be evaluated exactly and the worst-case finite-alphabet regression rate can be matched from above and below.

For the exact-risk statement, let x be uniform on $\{0,1\}^d$, and let a width-b interface copy the first b bits into the readout, giving $L_b = 2^b$ distinct patterns. For a square-integrable target $f(x)$ observed with independent zero-mean noise of variance σ^2, define $f_b(x) = \mathbb{E}[f(x) \mid x_1, \dots, x_b]$, $A_b = \mathbb{E}[(f - f_b)^2]$, and $M_b = \mathbb{E}[f_b^2]$. For solver ridge parameter $\alpha = m\gamma > 0$, let $N \sim \mathrm{Binomial}(m, 1/L_b)$ be the training count of an arbitrary pattern.

Proposition 9 (Exact approximation, occupancy, and estimation decomposition)**.** *The expected population excess risk of kernel ridge regression satisfies the exact decomposition*

$$\begin{aligned}\mathbb{E}[R(\hat{h}_{\alpha,b}) - R(f)] = A_b + M_b\,\mathbb{E}\frac{\alpha^2}{(N+\alpha)^2} \\ + (A_b + \sigma^2)\,\mathbb{E}\frac{N}{(N+\alpha)^2}.\end{aligned} \tag{108}$$

For the zero-ridge minimum-norm predictor ($\alpha \to 0^+$) predicting zero on unseen patterns, the exact expected excess risk and its non-asymptotic bound are

$$\begin{aligned}A_b + M_b(1 - L_b^{-1})^m + (A_b + \sigma^2)\mathbb{E}\left[\frac{\mathbf{1}\{N > 0\}}{N}\right] \\ \le A_b + M_b e^{-m/L_b} + \frac{2(A_b + \sigma^2)L_b}{m+1}.\end{aligned} \tag{109}$$

Neither identity requires a covariance-resolution event or a lower bound on the expected number of observations per pattern m/L_b.

Proof of Proposition 9. We first condition on a single pattern. For each pattern $s \in \{0,1\}^b$, let μ_s and v_s denote the conditional mean and conditional variance of $f(x)$ given $x_{1:b} = s$. Conditionally on the count $N_s = n$, the training label sum on pattern s has mean $n\mu_s$ and variance $n(v_s + \sigma^2)$, and KRR predicts this sum divided by $n + \alpha$. Its conditional expected squared error relative to μ_s is therefore

$$\frac{\alpha^2\mu_s^2 + n\,(v_s + \sigma^2)}{(n+\alpha)^2},$$

whose first numerator term is the shrinkage bias of the ridge and whose second is the averaged label noise.

Because the input distribution is uniform, every pattern shares the same marginal count distribution $N_s \sim \mathrm{Binomial}(m, 1/L_b)$, even though μ_s and v_s may vary across patterns. Averaging this expression over patterns therefore needs only the two pattern averages

$$L_b^{-1}\sum_s \mu_s^2 = M_b, \qquad L_b^{-1}\sum_s v_s = A_b,$$

while the orthogonal conditional-expectation decomposition adds the irreducible approximation error A_b. This establishes Eq. (108) without assuming homoscedastic conditional variance.

Taking $\alpha \to 0^+$ requires separating the event $n = 0$, on which the pattern is never observed, from the event $n \ge 1$. Separating these events before taking the limit yields Eq. (109). Its upper bound follows from the binomial identity $\mathbb{E}[(N+1)^{-1}] = \frac{L_b}{m+1}\left[1 - (1 - L_b^{-1})^{m+1}\right]$, together with $1/n \le 2/(n+1)$ for $n \ge 1$ and $(1 - L_b^{-1})^m \le e^{-m/L_b}$. □

For the minimax statement, fix $L = 2^b$ and consider the CNOT-copy circuit $|s\rangle_A|0\rangle_B \mapsto |s\rangle_A|s\rangle_B$ with s uniform over the L binary patterns. Its linear readout kernel is $k(s,t) = \mathbf{1}\{s = t\}$, so $D = L$. In the finite-alphabet regression model $y = \theta_s + \xi$, let $\xi \sim \mathcal{N}(0, \sigma^2)$ independently and

$\theta \in [-F, F]^L$, and write $\mathcal{M}_{m,L} = \inf_{\hat{f}} \sup_{\theta \in [-F,F]^L} \mathbb{E}_\theta \|\hat{f} - \theta\|^2_{L^2(P)}$, where the infimum is over all estimators based on m i.i.d. labelled observations.

Proposition 10 (Sharp learning cost of a saturated controlled interface)**.** *For every integer $m \geq 1$,*

$$\begin{aligned} \mathcal{M}_{m,L} &\geq \max\left\{F^2(1-L^{-1})^m, \frac{3}{128}\min\left(F^2, \frac{\sigma^2 L}{m}\right)\right\}, \\ \mathcal{M}_{m,L} &\leq \min\left\{F^2, F^2(1-L^{-1})^m + \frac{2\sigma^2 L}{m+1}\right\}. \end{aligned} \tag{110}$$

The second upper bound follows by bounding the risk of the zero-ridge minimum-norm kernel estimator, which averages labels within each observed pattern and predicts zero on unseen patterns. For a fixed constant $c_0 > 0$, if $F^2(1-L^{-1})^m \leq c_0\sigma^2 L/m$, these bounds give $\mathcal{M}_{m,L} = \Theta(\min\{F^2, \sigma^2 L/m\})$, with constants depending only on c_0. In the additional unsaturated regime $\sigma^2 L/m \leq F^2$, this rate reduces to the minimax order $\sigma^2 D/m = \sigma^2\, 2^b/m$.

Proof of Proposition 10. For the missing-pattern lower bound, put the uniform prior on the hypercube vertices $\theta \in \{-F, F\}^L$. If pattern s is unobserved, its sign is independent of the training data, so every estimator has a conditional Bayes squared error of at least F^2 on that pattern. Each pattern is unseen with probability $(1-1/L)^m$. Averaging over patterns proves the first lower bound, since a worst-case risk is at least any Bayes risk.

For the noise lower bound, use the hypercube $\theta \in \{-a, a\}^L$ with $a^2 = \min\{F^2, \sigma^2 L/(16m)\}$. The joint-data distributions at two vertices differing only at coordinate s have a Kullback–Leibler divergence of $2ma^2/(\sigma^2 L) \leq 1/8$: an observation reaches s with probability $1/L$, and the conditional Gaussian means differ by $2a$. Pinsker's inequality bounds their total variation by $1/4$. Any test for this sign therefore has an average error of at least $(1-1/4)/2 = 3/8$ under the equally weighted pair. An incorrectly signed estimate has a squared error of at least a^2. Pairing all hypercube vertices along each coordinate and averaging over the L coordinates gives a minimax risk of at least $3a^2/8 \geq (3/128)\min\{F^2, \sigma^2 L/m\}$.

For the upper bound, let N_s be the training count of pattern s. The conditional error of its sample mean is σ^2/N_s if $N_s > 0$, and the zero prediction has error θ_s^2 otherwise. With $N_s \sim \mathrm{Binomial}(m, 1/L)$, the inequality $1/n \leq 2/(n+1)$ for $n \geq 1$ and the binomial identity

$$\mathbb{E}\frac{1}{N_s+1} = \frac{L}{m+1}\left[1-(1-L^{-1})^{m+1}\right] \tag{111}$$

give the second term of the upper bound after averaging over s and using $\theta_s^2 \leq F^2$. The constant zero estimator has a risk of at most F^2, which gives the alternative upper bound. Under the stated missing-pattern condition, the upper bound is at most $\min\{F^2, (c_0+2)\sigma^2 L/m\} \leq (c_0+2)\min\{F^2, \sigma^2 L/m\}$, while the noise lower bound supplies the corresponding factor $3/128$. Removing the minimum additionally requires $\sigma^2 L/m \leq F^2$; missing-pattern domination alone does not imply this condition. □

F.3 Finite-Shot Spectral Resolution

For an acquired Gram matrix $\hat{K}$, define $D_{\mathrm{vis}}(\tau) = \#\{j : \lambda_j(\hat{K}) > \tau\}$. The following result relates this operationally visible dimension to the exact kernel spectrum under independent bounded per-entry measurements.

Proposition 11 (Finite-shot visible dimension)**.** *Suppose each entry K_{ij} $(i \leq j)$ of an $n \times n$ Gram matrix K is estimated by the sample mean of S independent random variables in $[-1, 1]$ with expectation K_{ij}, independent across unordered pairs, and let $\hat{K} = K + E$ denote the resulting symmetric estimator. With probability at least $1-\delta$, the operator error satisfies*

$$\|E\|_2 \leq \eta_S := C\left[\sqrt{\frac{n\log(2n/\delta)}{S}} + \frac{\log(2n/\delta)}{S}\right] \tag{112}$$

for a universal constant C. Under the controlled interface, for every resolution threshold $\tau \geq \eta_S$,

$$D_{\mathrm{vis}}(\tau) \leq \#\{j \leq D : \lambda_j(K) > \tau - \eta_S\} \leq D \leq 2^b. \tag{113}$$

Conversely, every exact eigenvalue satisfying $\lambda_j(K) > \tau + \eta_S$ satisfies $\lambda_j(\hat{K}) > \tau$ and remains visible.

Proof of Proposition 11. Write the symmetric error matrix as a sum of independent, centered self-adjoint matrices, one per shot and unordered pair. Each summand has an operator norm bounded by $O(1/S)$. For the variance parameter, the ith diagonal entry of $\sum_k \mathbb{E}X_k^2$ accumulates n pairs times S shots of $O(1/S^2)$, giving $\leq 4n/S$, and the variance matrix is diagonal, so it is $O(n/S)$. With these two parameters, the matrix Bernstein inequality (Tropp, 2015) establishes Eq. (112). Weyl's perturbation inequality gives $|\lambda_j(\hat{K}) - \lambda_j(K)| \leq \eta_S$ simultaneously for all $j = 1, \ldots, n$. Because $\lambda_j(K) = 0$ for all $j > D$, no such eigenvalue can exceed a threshold $\tau \geq \eta_S$; applying both sides of Weyl's inequality to the remaining indices proves both claims. Clipping the negative eigenvalues of $\hat{K}$ to zero leaves every positive eigenvalue, and hence the statement, unchanged. □

F.4 Interface-Sampling Complexity

When the controlled-interface branch geometry G is known, one can sample the control outcomes directly rather than estimate every kernel entry independently. The next two propositions give a dimension-independent upper bound for this estimator and a matching lower bound in its dependence on the number of inputs and target entrywise accuracy.

Proposition 12 (Dimension-independent interface sampling accuracy)**.** *For each of n inputs, draw S independent control-basis patterns from $p(x_i)$, form their empirical frequency vectors $\hat{p}_i$, and construct the factorized estimator $\hat{K} = \hat{Q}G\hat{Q}^\top$ using the exact branch-overlap matrix G. Then $\hat{K} \succeq 0$, $\mathrm{rank}(\hat{K}) \leq L$, and, with probability at least $1 - \delta$,*

$$\|\hat{K} - K\|_{\max} \leq \frac{2(1 + \sqrt{\log(n/\delta)})}{\sqrt{S}}. \tag{114}$$

Thus $O(n\varepsilon^{-2}\log(n/\delta))$ total control-state preparations suffice to achieve a maximum entrywise error of $\|\hat{K} - K\|_{\max} \leq \varepsilon$, with no dependence on the interface dimension L in this sampling bound. This sampling bound accounts strictly for state preparations and measurements; it excludes producing or calibrating G and the classical cost of assembling the output matrix.

Proof of Proposition 12. Define the empirical mean state $\hat{\rho}_i = S^{-1}\sum_{a=1}^{S} \tau_{s_{ia}}$, which serves as an analytical device and need not be explicitly constructed since $\mathrm{Tr}(\hat{\rho}_i\hat{\rho}_j) = \hat{p}_i^\top G \hat{p}_j$. By the independence of the S sampled patterns, the expected squared Hilbert–Schmidt error satisfies

$$\mathbb{E}\|\hat{\rho}_i - \rho_i\|_{\mathrm{HS}}^2 = \frac{\sum_s p_s(x_i)\,\mathrm{Tr}(\tau_s^2) - \mathrm{Tr}(\rho_i^2)}{S} \leq \frac{1}{S}, \tag{115}$$

where the inequality follows because $\mathrm{Tr}(\tau_s^2) \leq 1$ for any density matrix and $\mathrm{Tr}(\rho_i^2) \geq 0$.

To obtain a high-probability bound, replacing a single pattern sample changes the norm error by at most $\sqrt{2}/S$, because the Hilbert–Schmidt distance between any two density matrices is at most $\sqrt{2}$. Applying McDiarmid's bounded-differences inequality (McDiarmid, 1989) and Jensen's inequality $\mathbb{E}\|\hat{\rho}_i - \rho_i\|_{\mathrm{HS}} \leq (\mathbb{E}\|\hat{\rho}_i - \rho_i\|_{\mathrm{HS}}^2)^{1/2} \leq S^{-1/2}$ gives

$$\Pr\Big\{\|\hat{\rho}_i - \rho_i\|_{\mathrm{HS}} > S^{-1/2} + t\Big\} \leq e^{-St^2}. \tag{116}$$

Taking $t = \sqrt{\log(n/\delta)/S}$ and applying a union bound over all n inputs yields the bound simultaneously for all inputs with probability at least $1 - \delta$.

The state error then controls the kernel error directly. Expanding the inner-product difference and applying the Cauchy–Schwarz inequality gives

$$\begin{aligned} |\hat{K}_{ij} - K_{ij}| &= |\,\mathrm{Tr}(\hat{\rho}_i\hat{\rho}_j) - \mathrm{Tr}(\rho_i\rho_j)| \\ &\leq \|\hat{\rho}_i - \rho_i\|_{\mathrm{HS}} + \|\hat{\rho}_j - \rho_j\|_{\mathrm{HS}}, \end{aligned}$$

since both ρ_i and $\hat{\rho}_i$ have a Hilbert–Schmidt norm of at most one, proving Eq. (114). Moreover, $\hat{K}$ is positive semidefinite as the Gram matrix of $\{\hat{\rho}_i\}_{i=1}^n$, with rank at most L because all features lie in the operator span of the L branch states $\{\tau_s\}_{s=1}^L$. □

Proposition 13 (Interface-sampling lower bound). *Fix the access model of Proposition 12: each query to input x_i returns one independent control-basis pattern drawn from $p(x_i)$, while G is known exactly. There exists a two-branch controlled-interface family with orthogonal branch states such that, for any $0 < \varepsilon \le 1/16$, every estimator that returns the full $n \times n$ Gram matrix with $\|\widehat{K} - K\|_{\max} \le \varepsilon$ with constant success probability requires*

$$\Omega(n\varepsilon^{-2}) \tag{117}$$

total interface samples. Thus the $O(n\varepsilon^{-2}\log(n/\delta))$ upper bound in Proposition 12 is minimax-optimal in its n and ε dependence up to logarithmic confidence factors within this access model.

Proof of Proposition 13. It suffices to consider the two-branch subclass with orthogonal branch states, so $G = I_2$. Include a reference input x_0 with $p(x_0) = (1, 0)$. For each remaining input x_i, $i = 1, \dots, n-1$, choose a hidden sign $\theta_i \in \{-1, +1\}$ and set

$$p(x_i) = \left(\frac{1}{2} + 2\varepsilon\theta_i, \frac{1}{2} - 2\varepsilon\theta_i\right), \qquad 0 < \varepsilon \le \frac{1}{16}. \tag{118}$$

Then the reference kernel entries are $K_{i0} = p_1(x_i) = 1/2 + 2\varepsilon\theta_i$. Any estimator satisfying $\|\widehat{K} - K\|_{\max} \le \varepsilon$ therefore recovers every sign θ_i by thresholding $\widehat{K}_{i0}$ at $1/2$.

Under control-sample access, samples from input x_i are Bernoulli observations with parameter $1/2 + 2\varepsilon\theta_i$. The Kullback–Leibler divergence between the two one-sample distributions corresponding to $\theta_i = +1$ and $\theta_i = -1$ is $O(\varepsilon^2)$. The standard Assouad direct-sum argument for the $(n-1)$ independent signs (Tsybakov, 2009) therefore implies that recovering all signs with constant probability requires $\Omega((n-1)\varepsilon^{-2})$ total Bernoulli samples. The budget may also be allocated adaptively across inputs. Writing N_i for the random number of queries to x_i under a total budget $N = \sum_i N_i$, the chain rule for adaptively sampled experiments gives $\mathrm{KL} \le \mathbb{E}[N_i] \cdot O(\varepsilon^2)$ for two hypotheses differing only at coordinate i; the two-point argument then forces $\mathbb{E}[N_i] = \Omega(\varepsilon^{-2})$ for every i, and summing over i gives $\mathbb{E}[N] = \Omega(n\varepsilon^{-2})$. Since each sample is one interface-state preparation and measurement, the same lower bound applies to any Gram estimator in this oracle model. This proves the claimed $\Omega(n\varepsilon^{-2})$ preparation complexity. □

F.5 Post-Interface Noise

A fixed input-independent channel applied after the interface is a linear map on the readout operator space and therefore cannot create new feature directions.

Proposition 14 (Post-interface noise cannot increase capacity). *Let $\mathcal{N}$ be any fixed trace-preserving linear quantum channel applied to readout register B after the controlled interface, and define $\tilde{\rho}_B(x) = \mathcal{N}(\rho_B(x))$. Then the noisy feature dimension and empirical Gram rank satisfy*

$$\tilde{D} = \dim \operatorname{span}\{\tilde{\rho}_B(x) : x\} \le D \le 2^b, \qquad \operatorname{rank}(\tilde{K}) \le \operatorname{rank}(K). \tag{119}$$

Proof of Proposition 14. By the linearity of the channel, $\operatorname{span}\{\mathcal{N}(\rho_B(x)) : x\} = \mathcal{N}(\operatorname{span}\{\rho_B(x) : x\})$, whose linear dimension cannot exceed that of its input span. Applying the identical linear-mapping argument to a finite sample yields the Gram-matrix rank inequality. This guarantee applies to local or global depolarizing noise, fixed coherent readout distortions, and internal crosstalk within subsystem B occurring after the interface. □

G Experimental Setup

This appendix specifies the circuits, representations, numerical diagnostics, and kernel evaluator used by the fixed-representation studies of Sections H.1, H.2 and H.4. The exact-population and separator constructions are specified in Sections H.1 and H.2.

G.1 Circuit architecture

On qubit j, layer ℓ applies the following rotations in temporal order:

$$R_Y(\theta_j), \quad R_Y(w_{j,0}^{(\ell)}), \quad R_Z(w_{j,1}^{(\ell)}), \qquad \theta_j = j\text{th feature rescaled to } [0, \pi], \tag{120}$$

where the fixed weights $w_{j,i}^{(\ell)}$ are drawn from $[0, 2\pi)$. Both data paths apply affine rescaling to $[0, \pi]$ and then direct angle encoding without an additional prefactor (Shaydulin and Wild, 2022; Canatar et al., 2023). The synthetic map uses the fixed range $[0, 2]$, whereas the MNIST map uses the per-component minima and maxima from the fixed analysis pool. (We write j for the qubit index here; q denotes the causal-separator size in corollary 2.) An intra-A CNOT chain follows the single-qubit rotations. The entangler applies a fixed random single-qubit layer, b cross-register CNOTs with controls in A and targets in B, and a second fixed single-qubit layer. Each gate of those two layers is $R_Z(z)R_Y(y)$ with y and z drawn independently and uniformly from $[0, 2\pi)$ under a fixed seed; this is a uniform-angle law rather than the Haar measure on $\mathrm{SU}(2)$. The encoder uses two layers ($\ell \in \{1, 2\}$) throughout the synthetic and transductive capacity experiments, with its weights drawn from the same uniform-angle law under a separate fixed seed. By default, B contains no internal two-qubit gates.

The readout-side factors are product layers. In the default circuits described here, both V_B and W_B in Eq. (3) are products of single-qubit unitaries. Because V_B and all target-side controlled actions in this construction are product across the readout qubits, the branch states τ_s in corollary 1 are product states. Hence $\rho_B(x)$ is a state on the b_{tgt} target qubits tensored with a *fixed pure* product state on the remainder of B. Direct simulation confirms this, yielding a non-target reduced purity of 1 to machine precision. A ceiling indexed by readout-subsystem size, applied to that effective support rather than to n_B, already yields $4^{b_{\mathrm{tgt}}}$. For the default CNOT sweeps with distinct controls and targets, $b_{\mathrm{eff}} = b_{\mathrm{tgt}} = b$, so which-path cancellation tightens the ceiling from 4^b to 2^b. An entangling V_B removes this coincidence, giving a purity of 0.423 on $B \setminus T$ while preserving the measured rank of 2^b.

The *localized* wiring assigns control $i \bmod n_A$ and target $n_A + (i \bmod n_B)$ to crossing i, and uses a unidirectional chain within A. It carries the synthetic sweeps of Sections 4.1 and 4.2. The *distributed* wiring spaces controls and targets evenly across their registers and uses a bidirectional intra-A chain. It carries the image sweeps of Sections H.2 and H.4. All angles and data seeds are fixed, so repeated runs produce identical Gram matrices.

G.2 Representations and tasks

Synthetic XOR family. For XOR classification, the raw input is $x = [t_1, t_2, \epsilon_1, \ldots]$, where $t_i \sim \mathcal{U}[0, 2]$ and $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$. The label is $y = \mathbf{1}[(t_1 - 1)(t_2 - 1) > 0]$. Before angle encoding, each coordinate is clipped to $[0, 2]$ and affinely mapped to $[0, \pi]$. The lower clip is deliberate and not a tie-breaking detail: because ϵ_i is centered, clipping sends about half the padding coordinates to exactly 0, making the padding directions sparse rather than symmetric about their mean. In the hard-XOR setting, every coordinate is drawn independently from $\mathcal{U}[0, 2]$, but only the first two determine the label. In the multi-XOR setting (n_A even), all coordinates are informative and $y = \mathbf{1}[\sum_{k=1}^{n_A/2}(t_{2k-1} - 1)(t_{2k} - 1) > 0]$. For the label-independent rank sweeps, only this unlabelled input distribution is sampled. For XOR regression, $y_{\mathrm{clean}} = (t_1 - 1)(t_2 - 1)$ and $y_{\mathrm{noisy}} = y_{\mathrm{clean}} + 0.3\,\mathcal{N}(0, 1)$.

The MNIST 3-vs-5 representation. Binary labels are mapped to $\{-1, +1\}$ and raw images (LeCun et al., 1998) are projected onto 24 principal components (Jolliffe, 2002). PCA is fitted on the full class-pair pool before drawing the 400-point analysis pool. The extrema of this 400-point pool (comprising the 320-point training reservoir and the fixed 80-point test block) then determine the per-component affine scaling to $[0, \pi]$, without standardization. Although this feature pipeline is transductive—precluding an inductive interpretation of absolute test risks—the representation remains identical across every (b, m) cell in the capacity comparison. These 24 components are angle-encoded into $n_A = 24$ encoder qubits, paired with $n_B = 24$ readout qubits for a total register size of $N = 48$. Independent Gaussian noise with a standard deviation of 0.3 is added exclusively to the training labels, and the test MSE is evaluated against the ground-truth noiseless labels. The MNIST risk sweep evaluates $b \in \{0, \ldots, 8\}$ and $m \in \{20, 30, 45, 68, 100, 150, 225, 320\}$.

G.3 Learning protocol and numerical diagnostics

Kernel ridge regression is trained on each precomputed Gram matrix. The normalized objective in the theoretical formulation corresponds to $(K + m\gamma I)$, whereas the `scikit-learn` interface (Pe-

dregosa et al., 2011) uses the coefficient matrix $K+\alpha I$, yielding $\alpha = m\gamma$. The synthetic and MNIST (b, m) sweeps instead fix $\alpha = 10^{-6}$ across all cells, preventing condition-specific tuning from concealing the underlying capacity trend. This regularizer is small relative to the leading eigenvalues, though not necessarily relative to the compressed spectral tail. The noiseless-error experiment sets $m = 600$ with noiseless labels and $\alpha = 10^{-8}$. Its finite-sample test MSE includes estimation and regularization effects alongside the RKHS approximation error.

Our Gram-rank diagnostic counts eigenvalues exceeding the adaptive threshold

$$10 \max\{0, -\lambda_{\min}, \epsilon\lambda_1\},$$

where ϵ is the double-precision machine epsilon and the eigenvalues are computed after matrix symmetrization. Taking only the negative part of $\lambda_{\min}$ estimates the numerical positive-semidefiniteness violation floor. This prevents the false eigenvalue suppression that would result from applying $|\lambda_{\min}|$ to strictly positive-definite matrices. This matrix-dependent diagnostic operates independently of the theoretical ceiling and does not constitute a certified bound on tensor-network contraction errors. Empirically, the eigenspectra in Table 2 exhibit spectral gaps of at least five orders of magnitude at the reported ranks. Across the MNIST grid, the numerical rank estimate remains 2^b for any multiplier between 3 and 10^3 in place of 10. In contrast, fixed relative thresholds such as $10^{-9}\lambda_1$ yield 224 directions at $b = 8$, while $10^{-7}\lambda_1$ detects $30, 55, 101, 159$ directions for $b \in \{5, 6, 7, 8\}$. For the direct density-operator feature checks in Table 4, we compute the singular value decomposition of the explicit feature matrix with an absolute tolerance of 10^{-9}, rather than thresholding the Gram eigenvalues.

We evaluate kernel entries using the scalar tensor-network contraction introduced in Section 3.2. Each encoded state $|\Phi(x)\rangle$ is represented as a circuit tensor network (Markov and Shi, 2008). We optimize a single contraction path (Gray and Kourtis, 2021), reuse it across all sample pairs, and evaluate pairs in parallel across CPU workers. For general encoders, the cut width alone does not bound the classical contraction cost (Schuch et al., 2007; Haferkamp et al., 2020). The large-register sweeps reported here ran on a single multi-core CPU server with 64 workers and no GPUs. The largest sweep, covering the full MNIST (b, m) grid, required approximately two hours of wall-clock time.

Independent evaluator checks. We validate the four-copy scalar contraction against an explicit dense ρ_B reference wherever dense state simulation is feasible, covering $n_A = n_B \in \{3, 5, 8, 10\}$. For $n_A = n_B \in \{8, 10\}$, this check covers the grid $b \in \{0, 2, 4, 6, 8\}$. Because simultaneous transposition of all density matrices leaves their Hilbert–Schmidt Gram matrix unchanged, kernel-level comparisons alone cannot detect transposition errors. We therefore test the partial trace directly on a complex state whose reduced density matrix is not invariant under transposition. Across all tested cells, the scalar contraction and the dense reference agree to within 3.3×10^{-14}. The largest deviation occurs at $b = 0$, with agreement tightening as b increases (reaching 1.4×10^{-15} at $b = 8$ for $n_A = n_B = 10$). We also implement the algebraically independent factored evaluator of Eq. (6), where $p(x)$ is obtained by contracting the pre-interface control marginal and G is computed from the fixed product branch states. For the $b = 5$ configuration with a product V_B, distinct CNOT crossings, and an off-diagonal input pair, this factored evaluator matches the four-copy contraction to within 4.3×10^{-15} at $N = 20$ and 3.1×10^{-13} at $N = 1000$. This single-pair check at each size complements the broader dense-versus-scalar comparison.

Scalar contraction identity. Let $\Phi(x)_{a\beta}$ denote the joint state amplitude with the encoder multi-index a and the readout multi-index β. The four-copy tensor network evaluates the Hilbert–Schmidt kernel entry via the scalar contraction

$$k_B(x, x') = \sum_{a,a',\beta,\beta'} \Phi(x)_{a\beta} \overline{\Phi(x)}_{a\beta'} \Phi(x')_{a'\beta'} \overline{\Phi(x')}_{a'\beta}. \tag{121}$$

Here, the repeated encoder indices a and a' execute the partial traces over register A, while the cross-contracted readout indices β and β' evaluate the Hilbert–Schmidt inner product $\mathrm{Tr}[\rho_B(x)\rho_B(x')]$.

G.4 THE FOUR-COPY SCALAR EVALUATOR

Testing Theorem 1 on systems with hundreds of qubits requires an evaluation strategy that bypasses the explicit construction of $\rho_B(x)$, which becomes intractable as a $2^{n_B} \times 2^{n_B}$ matrix beyond a few dozen readout qubits. Instead, we contract four copies of the circuit directly to the scalar kernel entry $k_B(x, x')$ via Eq. (121), contracting register-A indices internally to trace out the encoder and cross-contracting register-B indices to evaluate the state overlap. Evaluating an $m \times m$ Gram matrix requires $O(m^2)$ such contractions, each governed by the contraction width and the total tensor count of the complete four-copy network (Markov and Shi, 2008; Vidal, 2003; Gray and Kourtis, 2021). Crucially, a bounded single-copy interface rank does not by itself bound this four-copy contraction cost. For the tested encoders, low-width contraction paths make direct evaluation of the exact kernel computationally efficient. The independent checks above assess its numerical accuracy. For circuit geometries with a large contraction width, Proposition 3 applies to controlled interfaces when control-basis samples are accessible and G is known or separately estimated (accounting for the calibration overhead). This replaces classical contraction by quantum sampling while retaining classical postprocessing in a dimension of at most 2^b.

H EXPERIMENTAL EVIDENCE

This appendix is organized by the claim each study tests rather than by experiment type: one subsection per question raised in Section 4, in the same order. Every study named here backs a specific sentence of the main text. The circuits, representations, and numerical diagnostics shared by all of them are specified in Appendix G.

H.1 ARE THE TWO CAPACITY CEILINGS ATTAINED?

Section 4.1 states that both ceilings are reached and that the gap between them is real. The following studies provide the evidence.

Two experiments test the structural ceiling of Theorem 1: numerical Gram rank against the crossing-gate family, and the sharper prefix-span prediction under local interleaving. Both measure the attained feature dimension; Section H.2 examines its statistical cost. Section H.1.3 is not an experiment: it records constructions that make each separator ceiling of corollary 2 attainable.

H.1.1 GRAM RANK UNDER CONTROLLED AND CROSSING-GATE INTERFACES

Table 2 reports the controlled-interface sweep and Table 3 the crossing-gate family referenced in Section 4.1.

Table 2: Numerical Gram rank under a controlled CNOT interface. Every reported entry equals the controlled-interface ceiling 2^b in corollary 1; dashes indicate $2^b > m$. For each reported rank, the retained eigenvalues are separated from the discarded numerical tail by a gap of at least five orders of magnitude.

b	0	1	2	3	4	5	6	7	8
ceiling 2^b	1	2	4	8	16	32	64	128	256
localized, $N \in \{20, 100, 200, 500, 1000\}$, $m = 40$	1	2	4	8	16	32	–	–	–
localized, $N = 100$, $m = 300$	1	2	4	8	16	32	64	128	256
distributed, $N = 48$, MNIST features, $m = 320$	1	2	4	8	16	32	64	128	256

Saturation comes from both sides of the interface factorization. For the largest controlled rank test ($N = 1000$, $b = 5$, and $m = 40$), we separately construct the control-marginal matrix Q and branch-overlap matrix G from Eq. (6). Both attain their maximum possible rank of 32, with a singular-value ratio of $\sigma_{\min}(Q)/\sigma_{\max}(Q) = 1.00 \times 10^{-2}$ and an eigenvalue ratio of $\lambda_{\min}(G)/\lambda_{\max}(G) = 8.35 \times 10^{-5}$. Their product satisfies $\mathrm{rank}(QGQ^\top) = 32$, with eigenvalues $\lambda_{32} = 1.88 \times 10^{-7}$ and $\lambda_{33} = 1.25 \times 10^{-15}$. The full column rank of Q and positive definiteness of G support saturation on the encoder and readout sides, respectively.

Table 3: Numerical Gram rank against the crossing-gate family at $n_A = n_B = 10$, $m = 300$, and a single circuit seed. The first two rows attain the common 2^b ceiling under both localized and distributed wirings. The Haar-random $b = 2$ entry and the last two rows were measured under the distributed wiring only. For the tested widths $b = 0, 1, 2$, the Haar-random $\mathrm{SU}(4)$ row has numerical Gram rank 4^b. The last two rows reverse the laboratory control direction (controlling from register B) but admit locally equivalent controlled decompositions, continuing to yield rank 2^b. Because $m = 300 > 4^4$, no entry is capped by the sample size. Dashes mark widths that were not evaluated.

Crossing gate	$b = 0$	1	2	3	4
CNOT	1	2	4	8	16
Controlled-R_y	1	2	4	8	16
Haar $\mathrm{SU}(4)$	1	4	16	–	–
reverse laboratory direction; locally equivalent controlled decompositions					
CNOT controlled from B	1	2	4	8	16
Controlled-R_y from B	1	2	4	8	16

Why reverse-control gates do not test the 4^b–2^b separation. The reverse-control CNOT and controlled-R_y gates in Table 3 admit locally equivalent controlled decompositions on A, with basis changes absorbed into the surrounding unitaries of Eq. (3). Their resulting Gram rank of 2^b is therefore consistent with the common-sector ceiling; the interleaving test below isolates whether rank-two crossings alone imply this ceiling.

Circuit-seed and spectral checks. The large-register and crossing-family grids use a single random draw of the encoder and entangler per configuration. To test sensitivity to random circuit realizations, a separate rerun fixes $N = 20$, $m = 40$, and a localized wiring, redrawing the encoder weights and the readout preparation across seeds $42, \ldots, 46$ for $b \in \{0, \ldots, 4\}$. All five seeds attain ranks $(1, 2, 4, 8, 16)$, with full eigenspectra archived alongside the integer counts. Across these 25 Gram matrices, the smallest spectral gap between the retained eigenvalues and the discarded numerical tail exceeds ten orders of magnitude. The shared-layout checks in Table 4 use fixed interfaces and a single seeded ensemble of input states per configuration, and the interleaving rows reproduce the reported feature ranks across five interleaving seeds with the sampled encoder states held fixed. Across all controlled configurations that saturate 2^b, the spectral gap between the retained eigenvalues and the discarded numerical tail spans five to fifteen orders of magnitude, including gaps of 12.9 and 10.4 orders for the multi-XOR spectra at $b = 3$ and $b = 5$, and a gap exceeding five orders for MNIST at $b = 8$. For the MNIST configuration at $b = 8$, the relative threshold $10^{-9}\lambda_1$, with $\lambda_1 \approx 6.6$, retains 224 eigenvalues, whereas the PSD-floor diagnostic retains 256. The relative threshold excludes weak eigenvalues that remain separated from the numerical tail. For every entry in Table 2, the ceiling-independent PSD-floor diagnostic from Appendix G falls within the observed spectral gap. All scalar Gram-rank scripts implement this adaptive threshold, while direct density-feature SVD checks use the absolute singular-value tolerance specified in Appendix G.

Rank-two interleaving can exceed the controlled ceiling. Table 4 inserts Haar-random unitaries on the control qubits between consecutive CNOT crossings, disrupting the common projectors $\{P_s\}$ while preserving an operator Schmidt rank of two at every crossing. The generic 4^b bound of Theorem 1 therefore still applies, but the common-sector argument need not. Without interleaving, all configurations attain their controlled ceilings; with interleaving, shared-control ranks rise from 2 to 4 for every tested $b \geq 2$. The distinct-control ranks reach 6 at $b = 2$ and 22 at $b = 3$, exceeding 2^b and attaining the prefix ceiling of Proposition 8. These sample ranks imply population dimensions of $D \geq 6$ and $D \geq 22$, respectively, showing that rank-two crossings alone do not imply the common-sector ceiling.

Sharing controls or targets cuts the observed span below 2^b. The one-to-one wiring in Table 2 has $b_{\mathrm{eff}} = b_{\mathrm{tgt}} = b$, whereas the six shared layouts in Table 4 reduce the CNOT counting ceiling by factors of two to eight. Exact dense simulation respects the refined ceiling $2^{\min\{b_{\mathrm{eff}}, b_{\mathrm{tgt}}\}}$ in all seven layouts and attains it in six. The exception has only two distinct branch states because both targets encode the same control parity, which falls below the counting ceiling of four. For the non-involutory target gate $R_y(0.7)$, repeated targets instead obey the common-unitary bound

Table 4: Layouts that change the effective interface at a fixed crossing count. **Top:** Controlled (ctrl.) and locally interleaved (int.) CNOT interfaces with the same rank-two crossings. Shared-control layouts reuse one control qubit, so their common-sector ceiling is $2^{b_{\mathrm{eff}}} = 2$. Exact dense simulation gives identical rank counts across five independent interleaving seeds; boldface marks ranks exceeding the corresponding non-interleaved ceiling. **Bottom:** Shared-control and shared-target Gram ranks from Haar-random encoder states and exact dense simulation at $n_A = n_B = 3$ (4 in the final two rows). The column $2^{\min}$ abbreviates $2^{\min\{b_{\mathrm{eff}}, b_{\mathrm{tgt}}\}}$; boldface marks CNOT saturation, and italics mark controlled-$R_y(0.7)$ ranks exceeding the CNOT count while respecting Eq. (15).

controls	b	n_A	$2^{b_{\mathrm{eff}}}$	2^b	4^b	rank(K), ctrl.	rank(K), int.
shared	1	3	2	2	4	2	2
shared	2	3	2	4	16	2	**4**
shared	3	3	2	8	64	2	**4**
shared	4	4	2	16	256	2	**4**
distinct	2	2	4	4	16	4	**6**
distinct	2	3	4	4	16	4	**6**
distinct	3	3	8	8	64	8	**22**
distinct	3	4	8	8	64	8	**22**

interface	b	b_{eff}	b_{tgt}	2^b	$2^{\min}$	rank(K) (CNOT)	rank(K) (c-R_y)
one per control, one per target	3	3	3	8	8	**8**	8
all crossings share one control	3	1	3	8	2	**2**	2
all crossings share one target	3	3	1	8	2	**2**	*3*
two controls, three targets	3	2	3	8	4	**4**	4
three controls, two targets	3	3	2	8	4	**4**	*6*
four crossings, two by two	4	2	2	16	4	2	3
five crossings, two controls	5	2	4	32	4	**4**	4

$\min\{\prod_\beta (m_\beta + 1), 2^{b_{\mathrm{eff}}}\}$ in Eq. (15). Two repeated-target layouts reach ranks of 3 and 6, exceeding the CNOT target-side counts of 2 and 4; in the other two, the unchanged control-side bound is already smaller. All controlled-rotation ranks satisfy the common-unitary bound and remain less than or equal to 2^b.

H.1.2 A PREFIX-SPAN BOUND FOR INTERLEAVED CONTROLLED CROSSINGS

Proposition 8, stated and proved in Section F.1, raises the generic 4^b ceiling to the prefix ceiling $(4^b + 2)/3$ once every crossing is binary A-controlled with arbitrary fixed local interleaving in between. The bound gives upper bounds of $2, 6, 22, 86$ on the rank at $b = 1, 2, 3, 4$, sharpening the interpretation of the distinct-control interleaving ranks 6 and 22 actually attained at $b = 2$ and $b = 3$ in Table 4. The prefix-span bound uses binary A-controlled crossings explicitly; Theorem 1 retains its operator-Schmidt statement for gates specified only by their rank.

We test both arbitrary local mixing and sector-preserving interleaving on $n_A = n_B = b$, with CNOT crossings and ten independently drawn local-unitary seeds per width. In the mixing arm, the interleaved encoder and readout unitaries are drawn from the Haar measure; in the preserving arm, the encoder unitaries are diagonal phases in the joint control basis, while the readout unitaries remain unrestricted. We use all d_A^2 states in a tomographically complete family, consisting of the computational-basis states and, for every $i < j$, the superpositions $(|i\rangle + |j\rangle)/\sqrt{2}$ and $(|i\rangle + i|j\rangle)/\sqrt{2}$. Their projectors span the full encoder operator space $\mathcal{B}(\mathcal{H}_A)$, removing uncertainty about whether randomly sampled inputs activate all available directions. In the mixing arm, the measured feature ranks reach $(4^b+2)/3$ across the tested widths, while in the sector-preserving arm they reach the 2^b ceiling. The full entanglers, input/output states, and singular spectra are archived. All ten seeds agree in every cell for both arms, showing numerical attainability through $b = 4$ without asserting an all-width saturation theorem.

H.1.3 QUANTUM SEPARATORS BEYOND CLASSICAL INTERFACES

Both the coherent quantum bound $D_R \le 4^q$ and the classical L-path bound in corollary 2 are attainable. For the quantum branch, the four single-qubit states with Bloch vectors $(1,1,1)$, $(1,-1,-1)$, $(-1,1,-1)$, and $(-1,-1,1)$ divided by $\sqrt{3}$ are linearly independent as density matrices. Their q-fold tensor products form an ensemble of 4^q input states whose density matrices span the full 4^q-dimensional operator space $\mathcal{B}(\mathcal{H}_C)$. An isometric suffix channel $\sigma \mapsto V\sigma V^\dagger$ into $n_R = q+2$ readout qubits is injective on operators and therefore achieves $D_R = 4^q$, despite the larger ambient readout space. Complete dephasing in the computational basis at the separator compresses this operator space to 2^q diagonal directions, while the CNOT-copy construction in Section H.2.3 attains the classical L-path bound with $L = 2^b$.

H.2 IS THE RISK RATE USABLE AT A FINITE BUDGET?

Section 4.1 states the constant in the $\sigma^2 D/m$ rate is close to one, so the rate sizes a budget rather than only ordering scalings. These studies compute the risk exactly and compare it against that scale.

These studies translate the attained feature dimension into prediction risk: an exact risk decomposition audited at every sample budget, a matching worst-case lower bound for the saturated controlled interface, and a kernel ridge study with an exactly known population covariance. Studies acting on the resulting width choice are in Section H.3, and the one that spends the budget on shots instead of labels is in Section H.4.

H.2.1 AN EXACT RISK DECOMPOSITION AT EVERY SAMPLE BUDGET

Let the input x be distributed uniformly on $\{0,1\}^d$, and let an interface of width $b \le d$ copy its first b bits into the readout register, inducing a nested categorical kernel with $L_b = 2^b$ distinct patterns. For any square-integrable target function $f(x)$ observed with independent zero-mean noise of variance σ^2, define the conditional expectation $f_b(x) = \mathbb{E}[f(x) \mid x_1, \dots, x_b]$, the approximation error $A_b = \mathbb{E}[(f - f_b)^2]$, and the accessible signal energy $M_b = \mathbb{E}[f_b^2]$. For a solver ridge parameter $\alpha = m\gamma > 0$, let $N \sim \mathrm{Binomial}(m, 1/L_b)$ denote the training sample count of an arbitrary pattern. The exact identity and its zero-ridge occupancy bound are stated and proved in Proposition 9 and Section F.2.

The exact decomposition shows that width selection depends jointly on approximation, pattern occupancy, and estimation variance. Widening the nested interface cannot increase A_b, but it enlarges the pattern space and reduces the expected number of observations per pattern. For independent bits and linear targets $f(x) = \sum_j a_j(2x_j - 1)$, the approximation error is explicit with $A_b = \sum_{j>b} a_j^2$, whereas a last-bit or full-parity target keeps $A_b = 1$ until the interface includes all required coordinates. When m/L_b is large and $\alpha \ll m/L_b$, the estimation term approaches $(A_b + \sigma^2)L_b/m$. The zero-ridge decomposition also applies when the sample-resolution condition required by the earlier high-probability KRR certificate is not met.

Exhaustive controlled experiment. We evaluate the exact risk decomposition across low- and high-occupancy regimes using eight-bit inputs ($d = 8$), without fitting scaling parameters. The grid covers interface widths $b \in \{0, \dots, 8\}$, sample budgets $m \in \{24, 48, 96, 192, 384, 768, 1536, 3072\}$, and noise levels $\sigma \in \{0, 0.1, 0.3\}$. The six pre-specified targets are normalized geometric sums with coefficients proportional to $2^{-\nu(j-1)}$ for $\nu \in \{0.5, 1, 2\}$, the first bit, the last bit, and full parity, with bit values in $\{-1, +1\}$. Each of the 1296 cells uses 2000 independent training/noise repetitions and minimum-norm zero-ridge fitting. Uniform fine-pattern counts and conditional Gaussian label sums are sampled exactly, then aggregated to each width; risks are evaluated on the entire 256-point population. The archive retains every target, noise level, width, budget, count, Gaussian draw, and risk, including configurations where the target is inaccessible or widening does not improve prediction. Figure 2(b) plots the finite binomial sum itself for a representative slice, rather than fitting a free scaling law to the curves. For a fixed pattern s, the probability that it remains unobserved is $(1 - L_b^{-1})^m$, and its contribution to the missing-pattern term on that event is $f_b(s)^2/L_b$. When such omissions are rare, a finite set of Monte Carlo repetitions may under-represent their contribution, especially in noiseless cells. We therefore also report an unbiased control-variate calculation: we subtract the realized contribution $L_b^{-1} \sum_s f_b(s)^2 \mathbf{1}\{N_s = 0\}$ from

each raw risk and add its known expectation $M_b(1-L_b^{-1})^m$. All raw outcomes remain archived; the adjustment integrates only this analytically known rare-event term and does not discard repetitions. Across the complete grid, the control-variate adjusted means differ from the exact formula by at most 2.57 standard errors of the adjusted means, with a 10^{-12} allowance for deterministic roundoff.

H.2.2 An unavoidable learning cost within the controlled family

Fix $L = 2^b$ and consider the CNOT-copy circuit mapping $|s\rangle_A|0\rangle_B \mapsto |s\rangle_A|s\rangle_B$, with inputs s drawn uniformly from L binary patterns. The induced linear readout kernel is $k(s,t) = \mathbf{1}\{s=t\}$, so $D = L$ and every regression function on these patterns belongs to its RKHS. Consider the finite-alphabet regression model $y = \theta_s + \xi$, with independent Gaussian noise $\xi \sim \mathcal{N}(0,\sigma^2)$ and a regression vector $\theta \in [-F,F]^L$, where $F, \sigma > 0$. Write $\mathcal{M}_{m,L} = \inf_{\hat{f}} \sup_{\theta\in[-F,F]^L} \mathbb{E}_\theta \|\hat{f} - \theta\|^2_{L^2(P)}$, where the infimum is taken over all estimators based on m i.i.d. labelled observations, not only KRR. The matching upper and lower bounds are stated and proved in Proposition 10 and Section F.2.

This is a worst-case result within one saturated, realizable controlled family, not a lower bound for every encoder or every target. The regression-function bound F is fixed across interface widths, while the largest RKHS norm in this target family is $F\sqrt{L}$. Thus the result establishes that the dimension-dependent upper bound cannot be uniformly improved over this uniformly bounded regression-function class, without claiming the same lower bound for a width-independent RKHS-norm ball.

H.2.3 KRR with an exactly known covariance

The controlled constructions in this and the following subsections specify the population feature distribution explicitly, allowing direct evaluation of covariance resolution, target approximation error, and test prediction risk. We fix all experimental configurations prior to sampling; each evaluation cell uses 2,000 independent training-data and Gaussian-noise realizations. The accompanying script `exp_theory_matched.py` records the random draws, per-replicate evaluation risks, and numerical verification checks against dense KRR solves.

Let the input string s be drawn uniformly from $\{0,1\}^b$, prepare the joint state $|s\rangle_A|0\rangle_B$, and apply a CNOT gate from each encoder qubit in A to its corresponding readout qubit in B. The resulting reduced feature state is $\rho_B(s) = |s\rangle\langle s|$, inducing the linear Hilbert–Schmidt kernel $k(s,t) = \mathbf{1}\{s=t\}$. This construction gives $D = 2^b$, $\kappa^2 = 1$, and the population covariance operator $\Sigma = I_D/D$, setting the exact spectral floor $\mu = 1/D$. The target function is the parity teacher $f^\circ(s) = (-1)^{\sum_j s_j}$, which belongs to this RKHS with squared norm $\|f^\circ\|^2_{\mathcal{H}} = D$ and is observed with independent $\mathcal{N}(0,0.3^2)$ label noise. For each interface width $b \in \{1,\dots,6\}$, the sample budgets and regularization parameters are set to

$$m \in \{m_0, 2m_0, 4m_0\}, \qquad m_0 = \lceil 20D\log(2D/0.05)\rceil,$$
$$\gamma = \frac{1}{mD}, \qquad \alpha = m\gamma = \frac{1}{D}. \tag{122}$$

Every evaluated configuration therefore satisfies the explicit sample-resolution condition established in the conditional-noise proof of Proposition 5. As b increases from 1 to 6, the baseline sample budgets m_0 evaluate to 176, 407, 923, 2068, 4579, and 10046.

Rather than sampling individual observation pairs, we simulate the sufficient statistics of the full i.i.d. experiment exactly: the category counts $(n_s)_{s\in\{0,1\}^b}$ follow a multinomial distribution, and conditional on the count n_s, the category label sum is $n_s f^\circ(s) + \sigma\sqrt{n_s}Z_s$ with independent standard normal variables $Z_s \sim \mathcal{N}(0,1)$. For any query point in category s, the resulting KRR predictor evaluates to this label sum divided by $n_s + \alpha$. The count-based predictions agree with an independent dense KRR solve, with a maximum absolute discrepancy of approximately 4.5×10^{-15} in the recorded numerical check. The population excess risk is computed by averaging the squared prediction error over all D discrete categories, avoiding finite test-set sampling approximations. Conditional on the category counts (n_s), its expectation is given analytically by

$$\frac{1}{D}\sum_s \frac{\alpha^2 f^\circ(s)^2 + \sigma^2 n_s}{(n_s+\alpha)^2}, \tag{123}$$

providing an independent consistency check on the empirical Monte Carlo estimates.

Across all 18 parameter configurations, the empirical mean excess risk normalized by the theoretical scale $\sigma^2 D/m$ ranges from approximately 0.989 to 1.027. All 36,000 generated realizations satisfy both the relative covariance event and the explicit excess-risk bound $4/m + 36\sigma^2(D + \log 40)/m$. The bound is conservative in this experiment, which tests the predicted risk scaling rather than the sharpness of the constants or coverage beyond the specified feature distribution.

H.3 Does target alignment decide what is learnable?

Section 4.2 examines how target information governs the benefit of widening and whether held-out selection identifies a useful interface from noisy labels. The following studies test these questions on quantum-generated data.

To test the task-alignment theory on genuinely quantum-generated inputs rather than on classical images, we report a controlled many-body experiment based on ground states of a disordered one-dimensional transverse-field Ising model (TFIM) (Pfeuty, 1970). The purpose is not to claim quantum advantage, but to isolate whether the interface exposes the quantum observable that defines the target.

Dataset and target. We generate 240 independent eight-qubit Hamiltonians

$$H(x) = -\sum_{j=0}^{6} J_j Z_j Z_{j+1} - \sum_{j=0}^{7} h_j X_j, \tag{124}$$

with $J_j \sim \mathrm{Unif}[0.65, 1.35]$ and $h_j \sim \mathrm{Unif}[0.35, 1.65]$, and compute their exact ground states $|\psi(x)\rangle$. The encoder is the abstract state-preparation map $U_A(x)|0\rangle = |\psi(x)\rangle$. The regression target is the long-range quantum correlation

$$y(x) = \langle\psi(x)|Z_0 Z_4|\psi(x)\rangle. \tag{125}$$

A CNOT interface copies selected computational-basis control bits to initially blank readout qubits, so the branch states are orthogonal and $G = I$. Consequently, if the selected controls include qubits 0 and 4, then $y(x)$ is exactly a linear function of the interface distribution $p(x)$:

$$y(x) = \sum_{s} (-1)^{s_0+s_4} p_s(x). \tag{126}$$

This makes the experiment a direct test of the target-projection statement in corollary 3.

Localized versus distributed interfaces. The localized order is $(0, 1, 2, 3, 4, 5)$, so both target qubits first become visible at width $b = 5$. The distributed order is $(0, 4, 2, 6, 1, 5)$, exposing both target qubits already at $b = 2$. For each width $b = 1, \ldots, 6$, we compute the exact interface probabilities from the ground-state amplitudes and the corresponding linear kernel $K = PP^\top$. We report two quantities: the empirical-population projection error obtained by least-squares projection of y onto the interface features using all 240 states, and KRR test MSE over ten fixed 160/80 train/test splits. The ridge parameter is selected by four-fold CV on each training split over $\{10^{-8}, \ldots, 10^{-1}\}$. As a classical sanity baseline related to kernel-based prediction of ground-state properties (Huang et al., 2022), we also fit an RBF kernel ridge model directly to the 15 Hamiltonian coefficients (J, h), tuning its bandwidth and ridge parameter within each training split.

The numerical transition is sharp. At $b = 2$, the distributed interface has projection error below 10^{-14} and mean test MSE 7.1×10^{-14}, whereas the localized interface has projection error 7.6×10^{-3} and test MSE 7.8×10^{-3}. The localized representation becomes exact only at $b = 5$, where its projection error falls below 2×10^{-14} and its mean test MSE to 8.0×10^{-9}. The RBF baseline on the Hamiltonian coefficients attains mean test MSE 6.81×10^{-4} (standard error 4.69×10^{-5}): it is better than under-aligned narrow interfaces but does not match an interface that directly contains the defining observable. These results give a quantum-generated counterpart to the classical-data experiments: the useful width is set by when the interface captures the target observable, not by the structural ceiling alone. Because the ground states are generated by exact diagonalization and the classical Hamiltonian parameters are available, this experiment is a target-alignment stress test rather than evidence of computational quantum advantage.

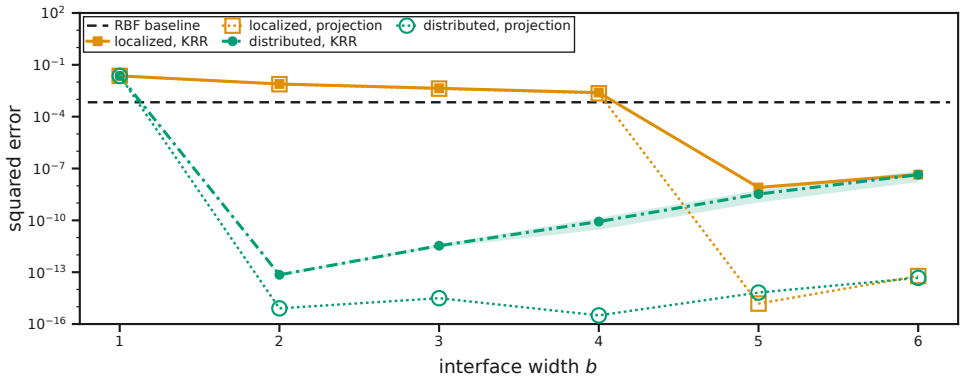


Figure 4: **Quantum-native target alignment on disordered TFIM ground states.** The target is the long-range correlation $\langle Z_0 Z_4 \rangle$. Solid and dash-dotted curves with filled markers are KRR test MSE (bands: ± 1.96 standard errors over ten splits); dotted curves with open markers are the exact finite-population projection error, which is the representability floor the risk can approach. The distributed interface exposes qubits 0 and 4 by $b = 2$, at which point the target becomes exactly representable; the localized interface requires $b = 5$. Increasing the width before the required target coordinates are exposed enlarges structural capacity without removing approximation error.

H.3.1 Hidden-Support Interface Selection

The preceding experiment intentionally exposes a known target support. To test whether the same interface principle remains useful when that support is not supplied to the model selector, we construct a second family of TFIM regression tasks on the same 240 ground states. Each task first samples four qubits and local X/Z readout bases. The clean target is a weighted sum of two disjoint two-body correlations on those four qubits,

$$y(x) = 0.6 \langle P_{q_1} P_{q_2} \rangle_x + 0.4 \langle P_{q_3} P_{q_4} \rangle_x, \tag{127}$$

where each pair uses a common X or Z basis so that the target is not forced to vanish by the TFIM parity symmetry. The support and bases are hidden from selection.

The candidate family contains every width-4 interface formed by choosing four of the eight qubits and assigning an X or Z readout basis to each, for a total of $\binom{8}{4}2^4 = 1120$ candidates. A basis rotation is followed by the same copy-to-readout construction used elsewhere, so each candidate exposes a 16-outcome joint distribution. For numerical convenience we use its 16 Walsh moments; this is an orthogonal rescaling of the joint probabilities and spans the same linear feature space. For each of 24 independently sampled targets, we use $m \in \{32, 64, 128\}$ training points, 56 validation points, and the remaining states for testing. Independent Gaussian noise with standard deviation 0.02 is added only to the training and validation labels. Ridge regularization is selected jointly with the interface using validation error over $\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$, and performance is evaluated against the clean target on held-out states. We compare the selected interface with an oracle that knows the target-defining interface and with a uniformly random width-4 interface using the same fitting protocol.

At $m = 32, 64, 128$, the selected interfaces achieve mean clean test MSE 1.38×10^{-4}, 7.06×10^{-5}, and 3.40×10^{-5}, respectively, compared with 1.26×10^{-4}, 6.0×10^{-5}, and 2.9×10^{-5} for the oracle. The corresponding random-interface errors are 3.49×10^{-3}, 2.96×10^{-3}, and 3.47×10^{-3}. Exact support-and-basis recovery is 91.7% at all three sample sizes. This experiment therefore tests a different question from the oracle-wiring audit above: the interface is selected from labeled data, and the relevant support is not provided to the selector in advance. The search space is still deliberately structured and small enough to enumerate, so the result should be read as evidence for data-driven interface selection rather than as a general architecture-search guarantee.

H.3.2 Interface topology, not width alone, sets the empirical minimum

Two experiments from the original fixed-representation study are reported here: a control showing that interface topology governs the empirical width trend, and a held-out width-selection experiment. Both use the transductive representation of Appendix G: $N = 48$ qubits split as $n_A = n_B = 24$, with images projected onto 24 principal components and angle-encoded. Across all cells, we fix the ridge regularizer to $\alpha = 10^{-6}$, add Gaussian noise of standard deviation 0.3 to the training labels only, and measure the test MSE against noiseless labels. These comparisons give within-protocol contrasts; absolute risks under this pipeline are not inductive estimates.

The empirical width ordering is a joint consequence of width and interface topology. Sweeping $b \in \{0, \dots, 8\}$ against $m \in \{20, 30, 45, 68, 100, 150, 225, 320\}$ under *distributed* wiring gives a risk-minimizing sequence $b^\star = (0, 0, 0, 3, 3, 3, 6, 8)$, whose Spearman rank correlation against $\log_2 m$ is $\rho = 0.95$ over all eight budgets and $\rho = 0.89$ over the five with $m \geq 68$. Evaluating the identical protocol under *localized* wiring instead gives $b^\star = (1, 8, 4, 4, 4, 5, 6, 6)$ and $\rho = 0.45$, exhibiting neither monotonicity nor a clear staircase progression. While the structural capacity ceiling 2^b depends solely on the number of crossings (Table 3) and holds under both wirings, interface width in isolation does not determine the empirical risk minimum.

H.3.3 Capacity-penalized width selection

This original paired experiment evaluates held-out width selection followed by refitting, choosing the interface width b without oracle access to the test risk. For each of 200 repeats and each total budget m, candidate KRR predictors are trained on $m_f = \mathrm{round}(2m/3)$ observations and evaluated on validation sets of size $m_v = m - m_f$ (for example, $m_f = 13$ and $m_v = 7$ at $m = 20$). The selectors minimize the validation MSE alone or the following penalized criterion:

$$\widehat{R}_{\mathrm{val}}(\hat{h}_b) + 0.09\,\frac{2^b}{m_f}. \tag{128}$$

The penalty coefficient 0.09 matches the injected label-noise variance ($\sigma^2 = 0.3^2 = 0.09$), specified prior to the comparison. The denominator is the candidate fit size m_f, while the coefficient is an empirical scale rather than a constant certified by a theorem. Because σ^2 is not known a priori outside this controlled setting, we also evaluate an adaptive coefficient obtained from the minimum validation risk, with no access to the injected noise level; the outcome of that variant is reported below. The predictor at the chosen width is subsequently refit on the full training subset of size m and evaluated on the fixed test block (200 points for synthetic regression and 80 points for MNIST). The oracle column records the lowest test risk among all candidate widths after refitting and serves solely as a reporting benchmark.

This experiment tests a practical sample-splitting and refitting workflow, whereas the Gaussian-label no-refit experiment in Section H.3.4 matches the extension in Section E.1. In this practical protocol, candidate predictions are unclipped, and post-selection refitting incorporates the validation observations. Consequently, the final refitted predictor is evaluated empirically rather than bounded directly by the fit-split selector guarantee.

The selection excess is the selected-and-refitted predictor's test MSE minus the minimum test MSE across all refitted candidate widths in the same repeat and budget. The capacity penalty reduces pooled selection excess by 30% on synthetic XOR regression and 39% on MNIST (Table 5). Seed-level paired Wilcoxon tests, evaluated on within-repeat averages across budgets, give $p = 0.0203$ on synthetic regression and $p = 7.76 \times 10^{-16}$ on MNIST. The per-budget columns reveal the empirical heterogeneity behind those averages. For synthetic regression, the penalty reduces selection excess by 53% and 40% at the two smallest budgets ($m = 20, 40$) but over-penalizes at larger sample budgets, most clearly at $m = 160$. For MNIST, it reduces selection excess by 42%, 58%, and 26% at the first three budgets ($m = 20, 40, 80$), performs comparably to unpenalized validation at $m = 160$, and gives an unresolved contrast at $m = 320$ ($p = 0.110$). The empirical evidence therefore supports the penalty as a scarce-data regularization policy in these tasks rather than a uniformly dominant selection rule.

The fixed design scale performs poorly as a selection rule in these experiments. Fixing $b = \lfloor \log_2 m \rfloor$ gives a pooled selection excess of 0.1417 on synthetic regression and 1.0978 on MNIST, exceeding

Table 5: Held-out width selection over $b = 0, \ldots, 8$ (mean test risk over 200 paired repeats). "Pen." adds the pre-specified $0.09\,2^b/m_f$ capacity penalty; $\log_2 m$ fixes the width at the design scale $\lfloor \log_2 m \rfloor$ instead of selecting it. Within each budget, p is the two-sided paired Wilcoxon signed-rank test (Wilcoxon, 1945) comparing Val. and Pen. across the 200 repeats. In the pooled excess row, test risks are first averaged across the five budgets within each repeat, and the Wilcoxon test is applied to those 200 seed-level paired means, avoiding pseudoreplication across nested budgets. All tests are conditional on the fixed representation, training reservoir, and test block.

	Synthetic XOR regression					MNIST 3-vs-5				
m	Oracle	Val.	Pen.	$\log_2 m$	p	Oracle	Val.	Pen.	$\log_2 m$	p
20	.1057	.1929	.1468	.4525	4.6×10^{-7}	1.0268	1.6736	1.4032	3.6041	4.4×10^{-10}
40	.0861	.1303	.1127	.2622	.053	.9882	1.3538	1.1421	2.6615	9.8×10^{-8}
80	.0572	.0842	.0855	.1357	.068	.9198	1.0558	1.0202	1.5702	.019
160	.0279	.0365	.0479	.0919	2.0×10^{-12}	.8276	.9542	.9547	1.3815	.429
320	.0156	.0161	.0171	.0585	.048	.7116	.7648	.7616	.7455	.110
Pooled excess		.0335	.0235	.1417	.0203		.2656	.1616	1.0978	7.76×10^{-16}

the unpenalized validation excess by factors of 4.2 and 4.1, respectively, while underperforming validation on 90.3% and 83.1% of paired budget–repeat cells (903/1000 and 831/1000).

Under an exponential approximation decay model, Eq. (55) gives the continuous minimizer of the approximation–estimation surrogate, which scales with a slope of $\log 2/(\nu + \log 2)$ against $\log_2 m$. Because finite-sample noiseless error fits do not identify the population approximation exponent ν, we do not evaluate heuristic plug-in estimates for that formula.

For a fully data-driven scale, we replace 0.09 with the minimum validation MSE within each repeat. Under this adaptive penalty, the pooled selection excess becomes 0.0253 on synthetic regression and 0.1155 on MNIST. On MNIST, this result represents a 57% reduction in selection excess relative to unpenalized validation (0.1155 versus 0.2656), improving upon the fixed-scale penalty (0.1616). On synthetic regression, however, the adaptive scale gives a higher pooled selection excess than the fixed penalty (0.0253 versus 0.0235). Because this scale is computed directly from the validation split, the resulting criterion is validation-dependent and falls outside the split-independent penalty condition of Proposition 4.

H.3.4 Gaussian-label selection with and without refitting

For inputs s drawn uniformly from $\{0,1\}^6$, an interface of width $b \in \{0, \ldots, 6\}$ copies the first b bits to the readout register via CNOT gates. These nested kernels induce exact population feature dimensions 2^b, with RKHS hypotheses spanning functions depending solely on the first b prefix bits. We define the target function $f^\circ(s) = \sum_{j=1}^{6} a_j(2s_j - 1)$ with normalized decaying coefficients $a_j = 2^{-(j-1)}/\sum_{k=1}^{6} 2^{-(k-1)}$. This target satisfies $|f^\circ(s)| \le 1$, and the target approximation error at width b is known analytically as $A_b = \sum_{j>b} a_j^2$. Labels are observed with additive independent Gaussian noise $\mathcal{N}(0, 0.3^2)$, each candidate predictor is fitted using a solver ridge parameter $\alpha = 10^{-6}$, and predictions are clipped to $[-1, 1]$. For total sample budgets $m \in \{24, 48, 96, 192, 384, 768\}$, we allocate $m_f = 2m/3$ observations to training and the remaining $m_v = m/3$ observations to independent validation. Model selection compares empirical validation risk alone against penalized validation using the pre-specified penalty $0.09\,2^b/m_f$ (Section E.1). The selected fit-split predictor is evaluated analytically over the full 64-point input population, while its separately refitted counterpart is retrained on all m observations at the selected width under the same clipping rule. Both evaluations are reported in Table 6.

At the two smallest budgets $m = 24$ and $m = 48$, capacity penalization lowers the mean no-refit risk by 0.00664 and 0.00248, with paired Monte Carlo standard errors of 0.00104 and 0.00050, respectively. At the largest budget $m = 768$, penalization instead slightly increases the no-refit risk by 0.000111 with a standard error of 0.000031. Refitting on all m samples reduces the mean risk across both selection rules at every evaluated budget, but alters the relative gap between selectors and falls outside the sample-splitting oracle guarantee. For the penalized no-refit selector, the Gaussian oracle inequality is satisfied in all stored realizations, although its remainder ranges from 6.10 to 1.08 and

Table 6: Mean population excess risk in the Gaussian-label selector experiment. Both columns within each evaluation pair share identical sample draws and noise realizations. The no-refit pair directly matches the theoretical framework of Section E.1, while the refit pair evaluates the practical heuristic of retraining on the full budget.

	No refit		Refit	
m	Validation	Penalized	Validation	Penalized
24	.07979	.07314	.07109	.06572
48	.04248	.04000	.03674	.03541
96	.02318	.02283	.01977	.02017
192	.01347	.01327	.01125	.01150
384	.00806	.00816	.00668	.00709
768	.00480	.00491	.00393	.00422

is too loose to explain the small observed selection differences. The substantive empirical finding is therefore the separately measured small-budget benefit and larger-budget over-penalization, rather than an assertion that the oracle inequality proves an improvement over pure validation.

H.4 What does a finite measurement budget cost?

Section 4.3 states that a finite shot budget resolves far fewer directions than the structural rank, that a fixed preparation budget has an interior optimum, and that reusable branch geometry makes the kernel cheaper to measure. These studies price each of those.

Finite-shot estimation alters the spectrum and learning behavior of linear projected kernels, and the interface representation governs the cost of acquiring the Gram matrix. Section H.4.3 reports the matched-budget comparison summarized in Table 1, Section H.4.5 repeats it on a circuit-derived configuration, and Section H.4.6 removes the analytically known-G assumption in a controlled measurement-model simulation.

Estimator. We model destructive SWAP-test estimation (Buhrman et al., 2001; Garcia-Escartin and Chamorro-Posada, 2013) on the readout register with independent per-pair measurement errors, matching the setting of Proposition 11. Randomized-measurement estimators that reuse snapshots across pairs can induce correlated errors (Huang et al., 2020; Elben et al., 2023). Each measurement shot returns a binary parity outcome $Y \in \{-1, +1\}$ with expectation $k = k_B(x, x')$, so the simulation draws the exact bounded-outcome estimator $\hat{k} = 2\,\mathrm{Binomial}(S, (1+k)/2)/S - 1$ rather than a Gaussian approximation. An independent experiment is executed for each unordered pair (i, j) with $i \le j$, giving a symmetric error matrix; diagonal entries are estimated on the same footing because the state purity $\mathrm{Tr}[\rho_B(x)^2]$ must likewise be measured on hardware. Because finite-shot sampling error can produce negative eigenvalues, KRR is evaluated on the PSD projection of the full 400×400 acquired Gram matrix under a protocol fixed prior to the experiments. The spectral diagnostics in Section H.4.1 instead evaluate the unprojected 300×300 principal training block. While projecting that block in isolation would retain its positive eigenvalues, principal submatrix extraction does not commute with full-matrix PSD projection; we therefore distinguish the raw training spectrum from the projected surrogate.

All other hyperparameters match the MNIST representation of Appendix G: $N = 48$, $n_A = n_B = 24$, distributed wiring, ridge regularizer $\alpha = 10^{-6}$, additive Gaussian training-label noise with standard deviation 0.3, 20 training subsets crossed with 10 label-noise realizations, a fixed 80-point test block, interface widths $b \in \{0, \ldots, 8\}$, and training budgets $m \in \{20, 30, 45, 68, 100, 150, 225, 320\}$. We generate 50 independent acquisitions for each per-entry shot budget $S \in \{10^3, 10^4, 10^5\}$, reusing each measured Gram matrix across its downstream training subsets as would be done in a physical experiment. Because shot budgets are specified per kernel entry, evaluating the full 400×400 data pool at $S = 10^3$ requires $80{,}200 \times 10^3 \approx 8 \times 10^7$ total shots for one width and one acquisition; we report S directly because it governs the entrywise error variance.

For each acquired matrix, let $D_{\mathrm{vis}}(\tau) = \#\{j : \lambda_j(\hat{K}) > \tau\}$ count the empirical eigenvalues that exceed a detection threshold τ. Proposition 11, stated and proved in Section F.3, connects this visible dimension to the exact spectrum through a high-probability matrix-Bernstein perturbation bound.

H.4.1 Shot budget sets the visible capacity

A measured Gram matrix has no stable exact-rank diagnostic: sampling can lift null eigenvalues away from zero, so the quantity bounded in Lemma 1 is not directly observable. In place of an exact rank, we employ two complementary resolution thresholds corresponding to two distinct operational questions. First, for a rough shot-budget estimate, we use the spectral-edge heuristic $2\sqrt{n/S}$ based on the worst-case entrywise variance bound $1/S$. Comparing an exact eigenvalue λ_k with this scale gives the heuristic budget condition

$$\lambda_k > 2\sqrt{n/S}, \qquad \text{that is,} \qquad S > 4n/\lambda_k^2. \tag{129}$$

Because this relation inverts analytically without requiring numerical simulation, we report the resulting estimates directly in Table 7. Second, to count the resolved directions in an empirically measured Gram matrix, our repeated-acquisition experiment fixes a two-sided matrix Bernstein threshold at the confidence level $1 - \delta = 0.95$ prior to sampling:

$$\eta_{S,\delta} = \sqrt{\frac{2n\log(2n/\delta)}{S}} + \frac{4\log(2n/\delta)}{3S}. \tag{130}$$

For the direct ± 1 parity estimator, each symmetric summand has an operator norm of at most $2/S$, and the sum has a matrix variance parameter of at most n/S, making Eq. (130) a concrete nonasymptotic instance of Eq. (112). With the threshold $\tau = \eta_{S,\delta}$, the condition $\lambda_j(K) > 2\eta_{S,\delta}$ is sufficient to guarantee visibility on the concentration event in Proposition 11. The visible-rank rule counts the eigenvalues of the measured 300×300 block that exceed $\eta_{S,0.05}$, independently of b, the exact spectrum, or the theoretical ceiling. The realized operator error was below the prespecified threshold in all 1350 simulated acquisitions.

On the 300-point training block of the MNIST configuration, a single eigenvalue dominates the Gram matrix at every width: λ_1 carries 85% of the trace at $b = 8$ and more at every narrower width, exceeding λ_2 by a factor of at least 21 for all $b \geq 1$. The spectral-edge heuristic gives an estimated requirement of 3×10^2 to 1.4×10^4 shots per entry for the second eigenvalue across $b \in \{1, \ldots, 8\}$; for the eigenvalue at the ceiling index, this requirement is 3.5×10^8 at $b = 2$, 7.2×10^{11} at $b = 3$, 1.5×10^{18} at $b = 6$, and 1.6×10^{24} at $b = 8$. At larger widths, the visible dimension $D_{\mathrm{vis}}(\tau)$ from Eq. (113) remains far below the structural ceiling 2^b within the tested shot budgets $S \leq 10^5$. For a fixed matrix size, increasing a target eigenvalue by a factor of $c > 1$ reduces its heuristic shot-budget estimate by a factor of c^2, without changing the structural ceiling at fixed b.

Table 7: Spectral-edge heuristic estimates of the per-entry shot budget needed to lift an eigenvalue above the noise floor in Eq. (129), on the 300-point MNIST training block. The leading eigenvalue λ_1 dominates across widths; at $b = 0$, the kernel is rank one with $\lambda_2 = 0$, so a budget for a second direction is not applicable. The continuous formula gives $S = 1.3 \times 10^{-2}$ at $b = 0$ as a formal baseline scale rather than an executable measurement count. The rightmost column estimates the cost of resolving the weakest direction when the controlled-interface ceiling $D = 2^b$ is saturated.

b	2^b	λ_1	λ_{2^b}	S for λ_2	S for λ_{2^b}
0	1	300.0	3.0×10^2	–	1.3×10^{-2}
1	2	152.0	9.5×10^{-1}	1.3×10^3	1.3×10^3
2	4	150.6	1.8×10^{-3}	1.4×10^3	3.5×10^8
3	8	85.2	4.1×10^{-5}	3.0×10^2	7.2×10^{11}
4	16	65.7	2.3×10^{-5}	3.2×10^2	2.4×10^{12}
5	32	36.5	5.2×10^{-7}	1.2×10^3	4.4×10^{15}
6	64	12.5	2.8×10^{-8}	5.9×10^3	1.5×10^{18}
7	128	16.9	4.6×10^{-8}	3.0×10^3	5.6×10^{17}
8	256	6.2	2.8×10^{-11}	1.4×10^4	1.6×10^{24}

H.4.2 Repeated direct-outcome acquisitions

Prediction risk depends on regularization and target alignment as well as on the kernel spectrum (Canatar et al., 2021). We therefore evaluate downstream KRR performance separately from the visible-dimension diagnostic. For every independent acquisition, we rerun all widths and training budgets with the same subsets and label-noise draws, changing only the measured Gram matrix. The exact positive-width sequence $(1, 1, 1, 3, 3, 3, 6, 8)$ has a Spearman correlation of 0.951 with $\log_2 m$. At $S = 10^3$, the acquisition-level median correlation is 0.000 with a 5–95% range of $[-0.609, 0.872]$ over the 43 nonconstant sequences, and 21 of all 50 acquisitions have a positive correlation. At $S = 10^4$, the corresponding median and range are -0.507 and $[-0.845, 0.620]$ over 44 nonconstant sequences, with 14 of all 50 acquisitions positive. At $S = 10^5$, the median correlation rises to 0.577 with a range of $[-0.041, 0.873]$ over the 36 nonconstant sequences, and 34 of all 50 acquisitions are positive. Thus the median correlations are nonmonotonic across shot budgets and remain below the noiseless benchmark even at $S = 10^5$.

Across all 1350 acquisition–width matrices, the Bernstein-resolved counts do not exceed 2^b: width bounds the available directions, whereas the spectrum and the shot budget determine visibility.

H.4.3 Kernel estimation without full probability reconstruction

The distribution-level ℓ_1 estimation route in Proposition 3 reconstructs more information than the linear kernel requires. When the branch-overlap matrix $G_{st} = \mathrm{Tr}(\tau_s \tau_t)$ is known, controlling the Hilbert–Schmidt error of the empirical estimates of $\rho_i = \sum_s p_s(x_i)\tau_s$ suffices to estimate $K_{ij} = \mathrm{Tr}(\rho_i \rho_j)$. Accurate ℓ_1 reconstruction of the probability vectors $p(x_i)$ is unnecessary. Because the branch states τ_s may be mixed, this representation also applies when the interface is followed by a fixed readout channel. Propositions 12 and 13, stated and proved in Section F.4, show that $O(n\varepsilon^{-2}\log(n/\delta))$ interface samples suffice for uniform entrywise accuracy while a two-branch subclass requires $\Omega(n\varepsilon^{-2})$ samples.

With independent acquisitions across distinct inputs, the off-diagonal kernel estimates are unbiased. The diagonal bias is nonnegative and equals the right-hand side of Eq. (115), reducing to $(1 - K_{ii})/S$ for pure branch states. Keeping the empirical diagonal entries preserves this estimator's Gram representation and guarantees positive semidefiniteness without post-hoc projection. When G is estimated rather than known analytically, a separately controlled positive-semidefinite calibration adds at most $2L\varepsilon_G$ to Eq. (114) by the argument of Eq. (25). We emphasize that this sampling improvement concerns the absolute entrywise kernel accuracy $\|\hat{K} - K\|_{\max}$, rather than the relative matrix error, the spectral resolution of near-zero eigenvalues, or downstream prediction guarantees under arbitrary regularization.

Matched acquisition experiment. For $b = 2, 4, 6, 8$ and $n = 32, 128$, we simulate kernel acquisition with fixed input distributions per width and input size, so that the 200 repetitions capture acquisition noise rather than input resampling. We fix $G = \bigotimes_{j=1}^{b} \begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}$, realized by pure product readout states with $|\langle\theta_j|X|\theta_j\rangle|^2 = 1/4$. Each input distribution combines a point mass of weight 0.6 at a uniformly drawn pattern with an independently drawn symmetric Dirichlet probability vector of weight 0.4, with a concentration parameter of $1/2$ per pattern, and can be prepared as the encoder state $\sum_s \sqrt{p_s}|s\rangle$. Across 200 independent acquisitions per cell, we compare three kernel estimators: independent destructive-SWAP estimates, their PSD projection (Higham, 1988), and interface sampling. The PSD-projected SWAP estimator reuses the raw SWAP measurements and requires no additional state preparations. To enforce an identical quantum preparation budget, we assign $T = 1, 4, 16$ SWAP shots per unordered pair and $S = (n + 1)T$ control samples per input, so that both acquisition protocols use exactly $n(n + 1)T$ state preparations, including the two preparations per shot required by the SWAP test and every diagonal pair. All 24 cells retain the empirical probability counts and SWAP counts, permitting the exact reconstruction of all three Gram estimates. The stored diagnostics include the maximum entrywise, relative Frobenius, and operator-norm kernel errors, the squared Hilbert–Schmidt feature-mean errors, and the realized diagonal errors, together with the analytic expectations of the latter two quantities. Table 1 compares the interface estimator against the strongest structure-matched reusable baseline on the 19 cells of the structure-matched baseline comparison, while Table 8 reports all 24 cells of the present grid for the three estimators that share the known branch geometry, including the $b = 8$ cells that the

baseline comparison does not cover; the $N = 1000$ configuration in Table 10 provides a separate circuit-derived comparison. Half-widths of 1.96 Monte Carlo standard errors reach at most 1.7% of a cell mean, so the means are reported alone.

Table 8: **The full known-G matched-budget acquisition grid.** All 24 configurations (n, b, T). Every estimator in a row receives the same $n(n+1)T$ state preparations; interface sampling uses $S = (n+1)T$ control samples per input. Means over 200 independent acquisitions.

				mean relative Frobenius error		
n	b	T	preparations	raw SWAP	SWAP + PSD	interface
32	2	1	1,056	2.163	1.567	0.1064
32	2	4	4,224	1.082	0.7977	0.05287
32	2	16	16,896	0.5373	0.4044	0.02633
32	4	1	1,056	5.041	3.617	0.1662
32	4	4	4,224	2.519	1.831	0.08185
32	4	16	16,896	1.252	0.9439	0.04157
32	6	1	1,056	9.523	6.766	0.2211
32	6	4	4,224	4.757	3.413	0.1124
32	6	16	16,896	2.38	1.757	0.05581
32	8	1	1,056	12.61	8.944	0.256
32	8	4	4,224	6.289	4.482	0.127
32	8	16	16,896	3.145	2.313	0.0631
128	2	1	16,512	2.167	1.549	0.05268
128	2	4	66,048	1.084	0.7787	0.02635
128	2	16	264,192	0.5416	0.3915	0.01316
128	4	1	16,512	5.371	3.823	0.0811
128	4	4	66,048	2.686	1.929	0.04052
128	4	16	264,192	1.342	0.9782	0.02029
128	6	1	16,512	11.98	8.486	0.109
128	6	4	66,048	5.998	4.278	0.05453
128	6	16	264,192	2.992	2.161	0.02726
128	8	1	16,512	20.13	14.25	0.129
128	8	4	66,048	10.06	7.142	0.06463
128	8	16	264,192	5.031	3.608	0.03226

H.4.4 Structure-matched reusable measurement baselines

Independent SWAP tests do not reuse outcomes across kernel pairs. We therefore compare against two reusable estimators motivated by randomized-measurement and classical-shadow protocols (Elben et al., 2020; Huang et al., 2020; Elben et al., 2023), and give all baselines the same PSD and rank information as the interface estimator. For a b-qubit readout, Pauli orthogonality gives

$$\mathrm{Tr}(\rho_i\rho_j) = 2^{-b} \sum_{P\in\{I,X,Y,Z\}^{\otimes b}} r_i(P)r_j(P), \qquad r_i(P) = \mathrm{Tr}(P\rho_i). \tag{131}$$

At each uniformly random Pauli string P_t, we measure $M = 3$ independent copies of every input, obtaining $Y_{ita} \in \{-1, +1\}$ for $a = 1, 2, 3$. For $i \neq j$, the random-Pauli estimate averages $2^b Y_{ita} Y_{jtc}$ over all t, a, c; on the diagonal it averages only $a \neq c$. This U-statistic form is unbiased for every entry.

The second estimator applies the same independently chosen local Pauli bases to every input state. For two measured bit strings $s, s' \in \{0, 1\}^b$, define

$$h_b(s, s') = 2^b(-2)^{-d_{\mathrm{H}}(s,s')}. \tag{132}$$

Averaging $h_b(s_{ita}, s_{jtc})$ over all settings and copies gives the local randomized-measurement overlap estimator (Elben et al., 2020); again, diagonal terms use $a \neq c$. The same setting supplies every pair. Exact one- and two-qubit density-matrix enumeration and a separate 10^5-sample Monte Carlo test verify both normalizations before the formal grid.

We use the product-branch family of Section H.4.3, fixing the inputs independently for each (n, b) and holding them fixed across budgets and acquisitions. For $n \in \{32, 128\}$, $b \in \{2, 4, 6\}$, and

$T \in \{1, 4, 16\}$, interface sampling uses $S = (n+1)T$ samples per input. Each reusable estimator uses $S_{\mathrm{RM}} = (n+1)T/3$ settings and three copies per input and setting. Since both 33 and 129 are divisible by three, every protocol uses exactly $n(n+1)T$ state preparations. We additionally test $(n, b, T) = (32, 8, 4)$, giving 19 cells. Each cell has 100 acquisitions.

Raw SWAP, random-Pauli, and local randomized-measurement matrices are symmetrized, projected onto the PSD cone (Higham, 1988), and then truncated to rank $r = \min\{n, 2^b\}$. This is a fairness control, not a guarantee of lower error for each realization; the constraint is inactive when $2^b \geq n$. The interface estimator is already PSD with rank at most r.

At $T = 4$, the exact means and Monte Carlo standard errors are reported in Table 9. Local randomized measurements are the strongest baseline at $b = 2, 4$, whereas SWAP becomes competitive or stronger at larger width. The known-G interface estimator remains best in all 19 cells. Relative to the strongest structure-matched baseline in each cell, its error is lower by factors from 4.44 to 78.19; the median factor is 18.73.

Table 9: Mean relative Frobenius error at the matched $T = 4$ budget. Parentheses give Monte Carlo standard errors over 100 acquisitions. Every baseline is PSD projected and rank truncated to $\min\{n, 2^b\}$.

n	b	SWAP	Random Pauli	Local RM	Interface, known G
32	2	.590 (.003)	.558 (.011)	.282 (.003)	.0490 (.0005)
32	4	1.815 (.006)	3.688 (.035)	1.526 (.011)	.0815 (.0007)
32	6	3.535 (.013)	26.878 (.260)	7.256 (.057)	.1134 (.0011)
32	8	5.015 (.017)	149.797 (1.251)	26.780 (.206)	.1359 (.0016)
128	2	.308 (.001)	.255 (.007)	.117 (.001)	.0263 (.0001)
128	4	1.558 (.002)	1.818 (.010)	.708 (.001)	.0409 (.0002)
128	6	4.257 (.003)	16.008 (.030)	4.308 (.009)	.0545 (.0002)

This experiment supports a specialized acquisition advantage from known reusable branch geometry in the evaluated family. It does not establish dominance over every optimized shadow ensemble, whose performance depends on its measurement ensemble and state class. The script `scripts/measurement_baselines.py` writes repeat-level results to `measurement_baseline_comparison_repeats.csv` and summaries to `measurement_baseline_comparison.csv`.

H.4.5 Matched-budget interface estimation

The factorized interface representation provides an alternative to estimating every Gram matrix entry via independent pairwise measurements. We compare raw entrywise SWAP estimation, its PSD projection, and the factorized estimator $\hat{Q}G\hat{Q}^\top$ on the archived $N = 1000$, $b = 5$, 40-input circuit configuration. Here the product readout branch states permit G to be computed analytically, corresponding to the known-overlap setting rather than an empirical hardware calibration. Each of the $40 \times 41/2 = 820$ unordered SWAP pairs uses S measurements and two readout-state preparations per measurement. The same $1640S$ state preparations allow $41S$ independent control-basis samples per input for the interface estimator. We run 200 independent acquisitions at each budget and evaluate the relative Frobenius kernel error (Table 10). The two acquisition protocols differ in circuit depth and measurement operators; the matched resource is the total state-preparation count, not the physical hardware runtime.

Table 10: Mean relative Frobenius error at matched total state-preparation budgets, evaluated with analytically known interface metric G. Parentheses report Monte Carlo standard errors over 200 independent acquisitions. Both the PSD-projected entrywise and factorized estimators yield valid positive semidefinite kernels; the factorized estimator additionally guarantees rank at most 32.

Total preparations	Entrywise	Entrywise + PSD	Interface, known G
1.64×10^5	.52790 (.00091)	.38360 (.00090)	.007095 (.000041)
1.64×10^6	.16654 (.00029)	.12350 (.00029)	.002232 (.000013)
1.64×10^7	.05261 (.00009)	.04035 (.00009)	.000716 (.000004)

In this configuration, the factorized estimator reduces the kernel error while remaining positive semidefinite up to numerical roundoff, with the smallest observed eigenvalue -1.5×10^{-15}. The plug-in quadratic estimates of the diagonal entries have a finite-sample bias, which is included in the reported Frobenius errors. When G is unknown, measuring branch overlaps incurs the calibration overhead analyzed in the proof of Proposition 3, and the PSD projection guarantees Eq. (25). This experiment therefore isolates the structural advantage of reusing a known low-dimensional interface metric, rather than claiming a universal advantage over entrywise measurements across all architectures.

H.4.6 FINITE-SHOT CALIBRATION OF UNKNOWN BRANCH GEOMETRY

We complement the known-G comparisons with a separate calibration experiment. For $b \in \{2, 4, 6\}$ and $L = 2^b$, the branch metric is $G = \bigotimes_{j=1}^{b} \begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}$, corresponding to product branches with local squared overlap $1/4$. We fix $n = 32$ input distributions, each mixing a point mass of weight 0.6 with weight 0.4 times an independent symmetric $\mathrm{Dirichlet}(1/2)$ vector, and use $S = 2048$ control-basis samples per input. The exact target is $K = QGQ^\top$.

We compare two ways of learning the reusable branch geometry. The first is a *generic dense* calibration that estimates every unordered off-diagonal entry of G independently from bounded SWAP outcomes. With C shots per branch pair, it costs $L(L-1)C$ branch-state preparations before PSD projection. The second uses the known product form of this benchmark: it estimates only the b local overlap factors and reconstructs their Kronecker product, reducing the calibration cost to $2bC$ preparations. This structure-aware control is essential here, since the $O(L^2)$ dense accounting is not an intrinsic lower bound for a branch family whose geometry factorizes. We intentionally exploit only the product factorization and not the additional parameter tying in this synthetic instance, where the exact local overlaps are equal; stronger prior structure could reduce calibration further.

Table 11: **Unknown-G calibration at the common endpoint $C = 2048$.** End-to-end max-entry error of the factorized estimator when the branch geometry is calibrated generically over every unordered branch pair, against calibration from the b local overlap factors alone, with the known-G sampling floor for reference. Parentheses give Monte Carlo standard errors over 200 repeats; the preparation columns are the one-time reusable calibration cost.

	end-to-end max-entry error			calibration preparations		
b	dense	product-aware	known-G floor	dense	product-aware	saving
2	0.03306 (0.00054)	0.02981 (0.00059)	0.02142 (0.00028)	24,576	8,192	3×
4	0.03085 (0.00032)	0.02786 (0.00034)	0.02570 (0.00034)	491,520	16,384	30×
6	0.03115 (0.00034)	0.02951 (0.00041)	0.02911 (0.00038)	8,257,536	24,576	336×

Over $C \in [16, 2048]$ the dense calibration-only error decays with fitted log–log slopes -0.462, -0.483 and -0.471 at $b = 2, 4, 6$, matching the $C^{-1/2}$ Monte Carlo rate at every width; what separates the two calibration schemes is therefore how many quantities must be learned, not how fast each one is learned. At $C = 2048$ the calibration-only errors are 0.02665, 0.02662 and 0.02617 for the dense estimator and 0.02110, 0.01595 and 0.01405 for the structure-aware one at $b = 2, 4, 6$; the end-to-end errors, the known-G floors and the one-time costs are in Table 11.

For every dense estimate we also compute $\varepsilon = \max_i \|\widehat{p}_i - p_i\|_1$ and $\varepsilon_G = \|\widehat{G} - G\|_{\max}$. All 4800 trials satisfy $\|\widehat{Q}G_+\widehat{Q}^\top - K\|_{\max} \le 2\varepsilon + 2L\varepsilon_G$ from Eq. (25). Before projection, 31.9% of the clipped symmetric estimates have a negative eigenvalue below -10^{-12}; after projection, all violations are at floating-point scale. The experiment therefore validates the generic entrywise estimator and, separately, shows that calibration complexity should be stated in terms of the structure used to parameterize G. Repeat-level results are written to `unknown_g_calibration.csv` and `structured_g_calibration.csv`.

H.4.7 FULL ACQUISITION-RESOURCE ACCOUNTING AND CALIBRATION AMORTIZATION

State preparations alone do not determine physical runtime. Table 12 therefore separates circuit executions, basis settings, input-pair configurations, additional logical two-qubit gates beyond state

preparation, reusable calibration, classical processing, and structural information. Here q_B is the number of measured readout qubits, $M = 3$ is the copies per input and reusable setting, and S_{RP} and S_{RM} are the numbers of shared random settings. A destructive SWAP execution consumes two state preparations and, under the stated Bell-basis circuit model, uses q_B logical entangling gates. Local randomized measurements use only single-qubit basis rotations during acquisition. These are logical circuit counts, not native-device latency estimates.

Table 12: Acquisition resources under the simulated logical circuit model. The unknown-G calibration terms are one-time and reusable; their displayed entangling-gate counts assume destructive SWAP calibration, either over all branch pairs or over the local product factors available in the structured benchmark. "Pair reuse" denotes whether one acquired setting contributes to all kernel pairs, not whether the measurement basis itself is fixed.

Protocol	Preparations	Settings / pair configurations	Extra logical 2Q	Reuse	Prior
SWAP + PSD + rank	$n(n+1)T$	1 basis; $n(n+1)/2$ input pairs	$q_B n(n+1)T/2$	No	PSD + rank
Random Pauli + PSD + rank	MnS_{RP}	S_{RP} shared settings	0	Yes	PSD + rank
Local RM + PSD + rank	MnS_{RM}	S_{RM} shared settings	0	Yes	PSD + rank
Interface, known G	nS	1 interface basis	0	Yes	Known G
Interface, unknown G (dense)	$nS + L(L-1)C$	Interface plus all branch pairs	$q_B L(L-1)C/2$	Yes	Generic calibrated G
Interface, unknown G (product-aware)	$nS + 2bC$	Interface plus b local factors	bC	Yes	Product form of G

For known G, generic dense assembly of $QGQ^\top$ costs $O(nL^2 + n^2L)$ time and $O(nL + L^2 + n^2)$ memory including the output. Generic dense calibration of an unknown G additionally requires $O(L^3)$ time for PSD projection and $O(L^2)$ calibration storage. When G has known product form, only the local factors need to be calibrated and the dense PSD-projection overhead is avoidable. SWAP projection and rank truncation cost $O(n^3)$ time and $O(n^2)$ memory. The direct local-randomized implementation costs $O(S_{\mathrm{RM}}M^2bn^2 + n^3)$ time; faster structured implementations would change classical time, not the preparation counts in the table. The generated numerical accounting is in `measurement_resource_accounting.csv`.

For the generic dense calibration protocol in Section H.4.6, let $C = S = 2048$. Its one-time cost is $L(L-1)C$, whereas each kernel build costs nS recurring interface preparations. For one build the calibration fractions at $n = 32$ are 0.273, 0.882, and 0.992 for $b = 2, 4, 6$; even at $n = 1024$, the $b = 6$ fraction remains 0.797. At $n = 32$, reusing the same branch geometry for 1, 8, and 126 kernel builds, respectively, reduces this dense-calibration contribution to at most one half of total preparations. These amortization numbers are therefore a stress test for generic entrywise calibration, not an intrinsic cost law for the product-structured benchmark; the structure-aware control above has one-time cost $2bC$.

For completeness, `measurement_accuracy_matched_break_even.csv` fits log–log error slopes from $T \in \{1, 4, 16\}$ and projects the baseline budget required to match each unknown-G endpoint. Every endpoint comparison extrapolates beyond the measured range, in some cases by orders of magnitude. We therefore retain it as a sensitivity calculation and do not use those projected break-even values to support a manuscript claim.

H.4.8 Global depolarizing noise on the readout

A fixed, input-independent quantum channel applied to register B after the controlled interface cannot increase the feature dimension; the general statement is Proposition 14, proved in Section F.5.

Table 13: **Labeled examples and interface shots compete under a fixed training–acquisition budget.** At $\mathcal{B} = mS = 4096$ the exact-feature KRR risk falls monotonically in m, while the S-shot interface predictor becomes measurement limited and its risk is minimized at an interior split. Population excess risk; parentheses give Monte Carlo standard errors.

m	S	exact features	S-shot interface
16	256	0.0300 (0.00029)	0.0322 (0.00029)
32	128	0.0268 (0.00018)	0.0311 (0.00018)
64	64	0.0251 (0.00012)	0.0335 (0.00012)
128	32	0.0242 (0.00008)	0.0401 (0.00008)
256	16	0.0241 (0.00006)	0.0526 (0.00005)
512	8	0.0237 (0.00004)	0.0714 (0.00003)

For global depolarizing noise, the transformation of the kernel can be evaluated in closed form. Under the channel $\rho_B \mapsto (1-p)\rho_B + p\, I/d$ with Hilbert space dimension $d = 2^{n_B}$,

$$\begin{aligned} \mathrm{Tr}[\tilde{\rho}_B(x)\tilde{\rho}_B(x')] &= (1-p)^2\, k_B(x,x') + \frac{p(2-p)}{d}, \\ \tilde{K} &= (1-p)^2 K + \frac{p(2-p)}{d}\, J, \end{aligned} \tag{133}$$

where J is the all-ones matrix, using $\mathrm{Tr}[\rho_B] = 1$ and $\mathrm{Tr}[I] = d$. The transformation clarifies how global depolarization affects the feature dimension and KRR predictions. First, for $0 \le p < 1$, global depolarization is invertible as a linear map: it preserves the identity component and scales every traceless component by $1-p$. It therefore preserves the feature dimension and the exact Gram rank. At the boundary $p = 1$, the channel maps all states to the maximally mixed state I/d, reducing the feature dimension and the Gram rank to 1. Second, for $0 \le p < 1$, omitting the additive rank-one term leaves a pure scalar rescaling by $(1-p)^2$; kernel ridge regression on $((1-p)^2 K,\, \gamma)$ produces predictions identical to KRR on $(K,\, \gamma/(1-p)^2)$ when both the training and train–test Gram matrices are scaled. Thus regularizer retuning can compensate for the multiplicative component, whereas the additive shift J cannot generally be absorbed into the ridge penalty. Third, at $n_B = 24$, $p = 0.5$, and $m = 320$, the operator norm of the additive component is $mp(2-p)/2^{24} = 1.43 \times 10^{-5}$, which exceeds the default ridge regularizer $\alpha = 10^{-6}$; a small entrywise scale does not by itself justify omitting this term. We therefore empirically compare predictions under the complete update against those under scalar rescaling alone across all widths, budgets, training subsets, and label-noise draws at each $p > 0$. Across the entire grid, the maximum pointwise prediction discrepancy remains below 2.3×10^{-7} and the maximum cell-mean test MSE change remains below 7.3×10^{-9}, confirming that multiplicative rescaling numerically dominates the noise effect in this experiment; over $m = 20, \ldots, 320$ the sequence of validation-selected widths is identical to the noiseless one except at $m = 150$, $p = 0.5$. For centered kernels, the additive shift vanishes algebraically because $HJH = 0$ for the centering projector $H = I - \mathbf{1}\mathbf{1}^\top/m$, provided training and test centering are applied consistently. This ablation, and Proposition 14, concern channels applied to the readout register *after* the interface, which cannot raise the feature dimension. Noise acting *within* the encoder or the crossings is outside the interface model of Eq. (3), which assumes a unitary entangler: a mid-circuit channel raises the Kraus rank of the transfer map and so need not respect the common-sector cancellation that gives 2^b. Characterizing the interface dimension under gate-level noise (Heyraud et al., 2022; Thanasilp et al., 2024) is left open.

H.4.9 A FIXED ACQUISITION BUDGET CREATES A SAMPLE–SHOT TRADEOFF

We isolate the two-resource law in Proposition 7 and corollary 7 using the smallest nontrivial controlled interface. The branch states are orthogonal ($G = I_2$), x is uniform on $[-1, 1]$, and

$$p_1(x) = \frac{1 + 0.7x}{2}, \qquad p_0(x) = 1 - p_1(x), \qquad f^\circ(x) = p_1(x) - p_0(x) = 0.7x. \tag{134}$$

Labels are $y = f^\circ(x) + 0.15\xi$ with an independent Rademacher variable $\xi \in \{-1, +1\}$, so $|y| \le 0.85$. We use ridge regularization $\gamma = 0.05$ and fix the total training-acquisition budget to $\mathcal{B} = mS = 4096$ interface state preparations. The labeled sample size is varied over $m \in \{16, 32, 64, 128, 256, 512\}$, which sets $S \in \{256, 128, 64, 32, 16, 8\}$ shots per input. Each

point averages 2000 independent training realizations. Population test risk is integrated on a 4001-point uniform grid; for the acquired predictor, the expectation over fresh S-shot query noise is evaluated analytically rather than by an additional Monte Carlo layer.

Table 13 reports both curves. The exact-feature risk decreases from 0.0300 at $m = 16$ to 0.0237 at $m = 512$. Under the fixed quantum-preparation budget, however, the acquired-feature risk is nonmonotone: it reaches its minimum 0.0311 at $m = 32$, $S = 128$, then increases to 0.0714 at $m = 512$, $S = 8$. The experiment is not used to fit the constants in Eq. (89) or to claim that its bound-level optimizer is exact. It tests the qualitative consequence of the theorem: once total acquisition is constrained, increasing m reduces the statistical term but simultaneously reduces S and can increase the measurement term.