ORDER-RESOLVED INTERACTION ATTRIBUTION IN NOISY QUANTUM MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature-interaction explanations determine whether a model uses features jointly or independently. For a deployed quantum classifier, reliable attribution requires separating interactions learned from data from those induced by the circuit and accounting for hardware noise. We develop an interaction-attribution framework that addresses these effects through the encoding, entangler graph, and noise channel. The encoding fixes the trigonometric degree of the decision function, allowing low-order Sobol variances to be identified exactly from a fixed grid. The entangler graph determines where architecture-induced interactions can occur. For the circuits studied here, the pairwise reach is $2(L-1)$ when the entangler commutes with the readout, and a hardware experiment on `ibm_fez` confirms this prediction for a distance-four interaction. Noise has a separate effect: its mode-wise contraction does not imply contraction of interaction variances, which can increase when noise breaks cancellations. We derive the attenuation law that holds and find that attenuation increases with interaction order for depolarizing and amplitude-damping noise. Across twelve datasets, the completeness residual is biased downward in all 60 single-encoding runs, while interaction rankings remain stable with mean Spearman correlation above 0.997. At higher noise, the two-point correction reduces the residual error from 0.0334 ± 0.0087 without correction to 0.0018 ± 0.0011. These results provide a circuit-aware framework for attributing interaction structure in deployed quantum models. Code: `https://anonymous.4open.science/r/ORIAN`.

1 INTRODUCTION

Feature interactions characterize whether a model uses its inputs jointly or treats their effects independently and additively. This distinction has direct operational consequences. A classifier that combines two features responds differently to perturbations than one that adds their individual contributions. An auditor who cannot distinguish these behaviors therefore has limited ability to identify the model's mechanism or understand its failure modes. Classical interpretability offers several estimators for this purpose, including Integrated Hessians (Janizek et al., 2021), the Shapley interaction and Shapley-Taylor indices (Grabisch & Roubens, 1999; Sundararajan et al., 2020), and Sobol sensitivity analysis (Sobol', 1990; Hooker, 2007). For a quantum classifier, however, these approaches generally treat the model as a black box and do not exploit structural properties that are known in advance.

Two such properties arise directly from the model construction. First, the trigonometric degree of a quantum decision function in each input coordinate is finite and determined by how often that coordinate is encoded (Schuld et al., 2021; Gil Vidal & Theis, 2020; Pérez-Salinas et al., 2020). Consequently, a function of known degree can be identified from a finite set of evaluations rather than estimated through generic sampling. Second, even before training, a parameterized quantum circuit is restricted to a function family shaped partly by its architecture (Qi et al., 2026; Thanasilp et al., 2024). The circuit can therefore display interaction structure that is present before learning. A faithful attribution method should distinguish such architecture-induced structure from interactions learned from data. Failure to make this distinction can lead to explanations that attribute model artifacts to learned behavior, a problem documented in both classical and quantum settings (Lapuschkin et al., 2019; Montalbano & Banchi, 2025).

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.

The device contributes another source of structure through noise. For circuits whose gates carry independent parameters, local Pauli noise attenuates each Fourier mode by a factor that decreases exponentially with the Hamming weight of its frequency vector (Fontana et al., 2025; Angrisani et al., 2026; Martinez et al., 2025). Since an order-$|S|$ component of the functional-ANOVA decomposition is constructed from modes supported on S, this behavior might suggest that higher-order interactions should be suppressed more strongly. This conclusion, however, does not generally hold for data-driven circuits. Several Pauli paths may contribute to the same interaction and can partially or completely cancel one another. Unequal noise-induced attenuation can disrupt such cancellations and thereby increase an interaction variance. Previous work exploits this contraction to truncate high-weight modes and enable classical simulation of noisy circuits (Fontana et al., 2025; Angrisani et al., 2026; Schuster et al., 2025). Its consequences for interaction explanations, however, remain unexplored.

These observations motivate an interaction analysis that follows the structure of the circuit. The encoding determines the possible interaction orders and enables finite-grid identification, while the entangler graph determines the support of architecture-induced interactions. Local noise, in turn, affects how these interactions survive measurement. This separation allows architectural effects to be calibrated once per ansatz and noise effects to be analyzed independently.

Our contributions are as follows.

- Exact identification of Sobol variances up to order k for re-uploading circuits and fidelity kernels using a fixed grid determined by the encoding degree.
- Characterization of architecture-induced interaction support for different entangler types, with hardware verification of the predicted depth dependence on `ibm_fez`.
- An exact attenuation law for operator weight and a corresponding bound for interaction variances, showing that noise can increase individual interaction variances. For depolarizing and amplitude-damping noise, this leads to a downward completeness-residual bias that grows with circuit depth.
- A two-point estimator that removes this bias without a noiseless reference, together with the noise and calibration regimes in which it is preferable to zero-noise extrapolation.

We make no claim of quantum computational advantage. We focus on the reliability and cost of interaction attribution for the deployed model. The identification in Section 3 is not specific to quantum models, since any decision function with known finite frequency support is determined on the corresponding grid. The noise correction in Section 5, however, has no classical counterpart.

2 Related work

Interpretability for quantum models has so far focused primarily on main effects under a single encoding (Gil-Fuster et al., 2024b; Heese et al., 2025; Pira & Ferrie, 2024; Steinmüller et al., 2022; Power & Guha, 2024; Baughman et al., 2022), without reporting interaction order. The identification framework used here follows from a known degree bound rather than from quantumness: recovering Sobol indices from a band-limited expansion is a classical result (Cukier et al., 1973; Rabitz & Aliş, 1999; Sudret, 2008; Sobol', 1990; Hooker, 2007). Similarly, the exponential damping of Fourier modes with their Hamming weight under local noise is well established (Fontana et al., 2025; Angrisani et al., 2026; Martinez et al., 2025) and has been used to truncate high-weight modes and simulate noisy circuits classically. These results provide the theoretical basis for our analysis, with a detailed comparison to prior work given in Appendix M.

3 Background

A variational quantum circuit encodes $x \in \mathbb{R}^n$ through angle gates interleaved with trainable unitaries and evaluates $f(x) = \mathrm{Tr}[\rho(x,\theta)\mathcal{M}]$ for a fixed observable $\mathcal{M}$ (Cerezo et al., 2021; Benedetti et al., 2019). A fidelity-kernel model predicts $f(x) = \sum_j \alpha_j k(x_j, x)$ with $k(x,x') = |\langle\phi(x)|\phi(x')\rangle|^2$ (Schuld, 2021; Havlíček et al., 2019). Throughout, f is the real-valued decision function and $\mathcal{M}$ a sum of single-qubit terms.

Under a product law $P_X = \bigotimes_i P_i$, every $f \in L^2(P_X)$ admits a unique functional-ANOVA decomposition $f = \sum_{S \subseteq [n]} f_S$ into mutually orthogonal components (Sobol', 1990; Hooker, 2007). Writing $D_S = \mathrm{Var}(f_S)$ for the interaction variance of a feature subset, the order-k *completeness residual*

$$R_k \;=\; 1 - \frac{\sum_{|S| \le k} D_S}{\mathrm{Var}(f)} \;\in\; [0, 1] \tag{1}$$

is the fraction of decision variance left unexplained by interactions of order at most k, and attribution to a feature pair is D_{ij}. Each P_i is uniform on the encoded range, following the deployed angle convention. Preprocessing rank-transforms each feature to a uniform marginal, which fixes each P_i individually but leaves the joint law non-product; Appendix I compares the grid residual with the residual under each dataset's own law.

Conditional expectation under a product law is linear and acts coordinate-wise on a finite trigonometric sum, so the conditional mean $\mathbb{E}[f \mid x_S]$ is itself a trigonometric polynomial in x_S with the same per-coordinate degree, and is identified exactly on a tensor grid.

Theorem 1 (Exact identification)**.** *Let f be the decision function of a re-uploading circuit in which feature i is encoded m_i times, under a product input law and a one-local observable. Then f has degree at most m_i in coordinate i, and for every $S \subseteq [n]$ the Sobol variance D_S is a closed-form functional of the $\prod_{i \in S}(2m_i{+}1)$ conditional-mean evaluations of f on the tensor grid of $2m_i{+}1$ nodes per coordinate. The set $\{D_S : |S| \le k\}$ is therefore exact from $1 + \sum_{1 \le |S| \le k} \prod_{i \in S}(2m_i{+}1)$ evaluations, which is polynomial in n for fixed k and $\max_i m_i$. The same conclusion holds for a fidelity kernel whose angle feature map encodes each feature r times, with $m_i \equiv r$.*

The identification procedure operates directly on the Fourier coefficients rather than through quadrature. The corresponding weight matrix is diagonal only at degree one, so a weighted sum of squared coefficients does not give the correct Sobol variances at higher degrees. The degree is determined by the number of feature encodings, not by the modulus-square operation; hence, a single-encoding fidelity kernel does not have degree two. The computational cost is instead dominated by the residual R_k, whose denominator requires all $2^n - 1$ nonempty subsets. All reported values of R_k use the full product grid, with reconstruction error below 1.2×10^{-14}.

A classical surrogate provides a direct control for the identification procedure. The coefficient tensor in Theorem 1 is the band-limited least-squares interpolant of f on the same node grid. Consequently, exact Sobol analysis of a surrogate fitted to these queries should recover the same variance numerators, which is observed at matched budgets: the surrogate achieves $R^2 = 1.000000$ at every degree (Table 4 in Appendix C). Thus, the distinction lies in the identification guarantee rather than in the resulting estimator, since the same queries suffice for a classical surrogate to recover the variances. The noise correction in Section 5 is different: it requires the encoding multiplicity and device Pauli eigenvalues, information that is not contained in input-output pairs.

4 Architecture-induced interaction structure

Exact estimation does not by itself ensure faithful attribution. The architecture can induce interactions before training. For a ring entangler, an ensemble of 50 untrained circuits gives a mean pairwise variance of 1.8×10^{-4} for adjacent pairs versus 1.2×10^{-5} for distant pairs, a ratio of about fifteen. We therefore standardize each D_S by its mean and spread under the initialization distribution used for the deployed model. On a label-shuffled null, this reduces spurious recovery from 0.50 ± 0.35 with a naive mixed-derivative probe to 0.20 ± 0.27 (mean $\pm$ standard deviation over five seeds). The limitation is computational cost: the reference ensemble must be resampled for each architecture and feature count. The following approach removes this cost at inference time.

Proposition 2 (Light-cone support)**.** *Let the untrained circuit have L layers of nearest-neighbour entanglers on a graph G whose gates within a layer act on disjoint pairs or commute with one another, with single-qubit encodings, trainable single-qubit layers drawn from any input-independent distribution, and an observable that is a sum of single-qubit terms. Let $d_G(i, j)$ be the graph distance between the qubits carrying features i and j. Then*

$$d_G(i, j) > 2L \implies D_{ij}(f_{\Theta_0}) = 0 \quad \textit{for every } \Theta_0. \tag{2}$$

Within the light cone the prior is a functional of the ansatz specification alone, and is constant on graph-distance classes whenever the automorphism group of G acts transitively on pairs at fixed distance.

Both hypotheses of Proposition 2 are necessary. Each feature must lie within the backward light cone of a single-qubit term in the observable, with radius L. Two features in such a cone can be $2L$ apart, giving the pairwise reach. The within-layer condition ensures that each layer can expand the support by one step. On eight- and ten-qubit rings across four depths and five seeds, $\mathrm{IsingZZ}$ also commutes with the one-local Z readout, so its final entangling layer does not expand the observable support. The pairwise reach therefore tightens to $2(L-1)$, with no exceptions in 1460 measurements, compared with 95 for a reach of L. In contrast, $\mathrm{IsingXX}$ commutes within a layer but not with the readout and exactly saturates $2L$. A CNOT ladder violates the $2L$ bound, reaching a distance-four pair at $L=1$ because sequential gates sharing qubits can propagate support across the chain within a single layer. Using CNOT gates on disjoint pairs restores the bound.

Since the prior depends on the ansatz but not on labels or trained parameters, it can be calibrated once and reused across models with the same specification. Its calibration error decreases as $O(N_{\mathrm{cal}}^{-1/2})$ (Hoeffding, 1963). Haar and anisotropic initialization laws on the correct graph differ from a 5000-circuit reference prior by 11-18%, comparable to the 5-11% sampling error at 500 circuits, whereas a wrong-graph prior differs by 389-554%. At a threshold fixed on the matched null, false-positive rates are 6.2%, 7.7%, and 4.6% for the three same-graph laws, compared with 15.4% for the wrong graph, while the true-positive rate is 1.00 throughout. Against a sampled null, the prior reduces cost rather than improving accuracy, recovering the planted pair at 0.78 ± 0.04 versus 0.83 ± 0.04. Appendix L separates the effects of graph structure and initialization law.

Proposition 2 rules out interactions from the ansatz alone, before training. For the trained model, a complementary certificate is obtained from the Heisenberg-evolved observable, $M_{\mathrm{eff}} = W^\dagger \mathcal{M} W = \sum_P m_P P$. Define the *operator certificate* $\Omega_{i:j} = \sum_{P:P_i \neq I, , P_j \neq I} |m_P|^2$. If M_{eff} is additively separable, every straddling ANOVA component vanishes; hence $D_{ij} > 0$ implies $\Omega_{i:j} > 0$. Within the cone, where Proposition 2 is silent, $\Omega_{i:j}$ tracks the trained D_{ij}: across twelve datasets, the median of the per-dataset median Spearman correlations is 0.745 over 60 trained models, exceeding the pre-set threshold of 0.70, with eleven datasets clearing it individually. The certificate requires one Pauli decomposition and no untrained ensemble, but characterizes the trained model rather than the architecture.

4.1 Testing Interaction Reach on Hardware

The depth dependence in Equation (2) can be tested directly on hardware, at lower cost than estimating D_{ij}. If no Pauli string in M_{eff} acts on both features, then f contains no term that depends jointly on the two coordinates. The mixed second difference

$$\Delta_{ij} = f(+h, +h) - f(+h, -h) - f(-h, +h) + f(-h, -h)$$

is therefore zero for every setting of the remaining coordinates. Evaluating it requires only four circuits per slice, rather than a full ANOVA grid.

We test this prediction on a six-qubit linear chain by planting a strong interaction between a distance-four pair and varying the circuit depth. At $L=3$, the reach $2(L-1) = 4$ includes the pair, whereas a reach equal to the circuit depth excludes it. At $L=2$, both bounds exclude the pair. The predictions were registered before submission. The experiment uses 24 random slices, three seeds per depth, and 2048 shots on `ibm_fez` (IBM Quantum, 2026), transpiled to physical qubits 0-5 with exactly $2(n-1)L$ two-qubit gates and no routing overhead.

Figure 1 and Table 1 show that all three registered predictions hold. At $L=3$, every interval excludes zero, whereas at $L=2$ every interval contains zero. Held-out accuracy separates the two depths in the same direction, 0.702 versus 0.805, since the depth-two circuit cannot represent the planted interaction. Thus, the hardware measurement supports a reach of $2(L-1)$ rather than one equal to the circuit depth. The measured values are 30%-35% below the calibrated noise-model prediction.

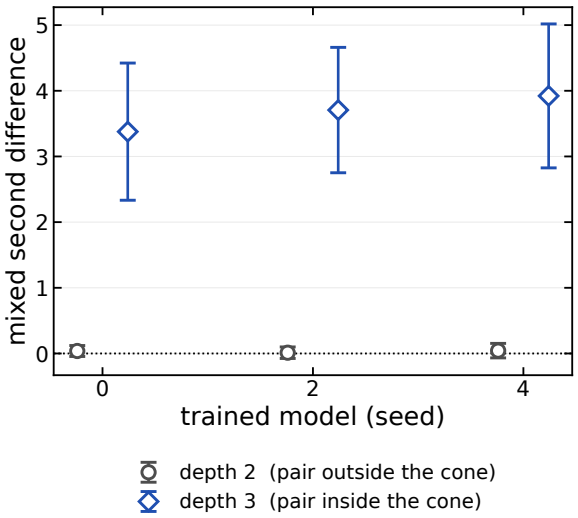


Figure 1: Hardware measurement of the planted distance-four interaction on `ibm_fez`.

Table 1: Hardware results for the planted distance-four pair.

Depth	Seed	Test accuracy	Mean Δ_{ij}	95% interval	Excludes 0
$L = 2$	0	0.711	$+0.038$	$[-0.046, +0.121]$	No
$L = 2$	2	0.707	$+0.012$	$[-0.076, +0.100]$	No
$L = 2$	4	0.689	$+0.043$	$[-0.067, +0.154]$	No
$L = 3$	0	0.814	$+3.378$	$[+2.333, +4.422]$	Yes
$L = 3$	2	0.811	$+3.706$	$[+2.752, +4.661]$	Yes
$L = 3$	4	0.791	$+3.922$	$[+2.826, +5.018]$	Yes

5 Order-selective attenuation under noise

The structural properties above describe the noiseless decision function. On hardware, the measured f also reflects the action of a noise channel, which can affect different interaction orders unequally.

For a circuit whose gates carry independent parameters, a local Pauli channel multiplies each Fourier coefficient by at most $q^{|\omega|}$ in the Hamming weight of its gate-level frequency (Fontana et al., 2025). This might suggest that D_S contracts at least as fast as $q^{2|S|}$. For data-driven circuits, however, this does not generally hold. Multiple Pauli strings can contribute to the same component f_S, including strings whose factors outside S contribute only a constant. These contributions can cancel in the noiseless circuit, while unequal attenuation can break the cancellation and increase the resulting interaction variance. The following theorem gives the corresponding bound.

Theorem 3 (Order-resolved attenuation)**.** *Consider a circuit in which feature k enters once, through $R_Y(\theta_k)$ on qubit k with θ_k uniform on $[0, \pi]$, followed by a trainable unitary W and an observable $\mathcal{M}$, and write $M_{\text{eff}} = W^\dagger \mathcal{M} W = \sum_P m_P P$. Let a local Pauli-diagonal channel with eigenvalues $\lambda_X, \lambda_Y, \lambda_Z \in (0, 1]$ act on every qubit immediately after the encoding layer. For $S \subseteq [n]$ let $Q_S = \sum_{\text{supp}(P)=S} m_P^2$ and $\lambda_S = \max_{a \in \{X,Z\}^S} \prod_{k \in S} \lambda_{a_k}$. Then: (a) $Q_S^{\mathcal{N}} = \sum_{\text{supp}(P)=S} \Lambda(P)^2 m_P^2$ with $\Lambda(P) = \prod_{k \in S} \lambda_{P_k}$, so $Q_S^{\mathcal{N}} = q^{2|S|} Q_S$ for an isotropic channel; (b) $D_S^{\mathcal{N}} \le \lambda_S^2 \bar{D}_S$, where $\bar{D}_S \ge D_S$ is the cancellation-free majorant of Equation (7); (c) if the contributions to each sector of S share a sign, $D_S^{\mathcal{N}} \le \lambda_S^2 D_S$, which for an isotropic channel reads $D_S^{\mathcal{N}} \le q^{2|S|} D_S$; (d) under an isotropic channel the operator-weight residual $R_k^Q = \sum_{|S|>k} Q_S / \sum_{S \neq \emptyset} Q_S$ strictly decreases whenever it is nonzero.*

Part (a) is exact and requires no sign condition, giving the guaranteed decrease of the operator-weight residual in (d). Part (c) recovers the order-$|S|$ contraction when the sector contributions do not cancel, while part (b) remains valid without that condition. The proof and sector definition are given in

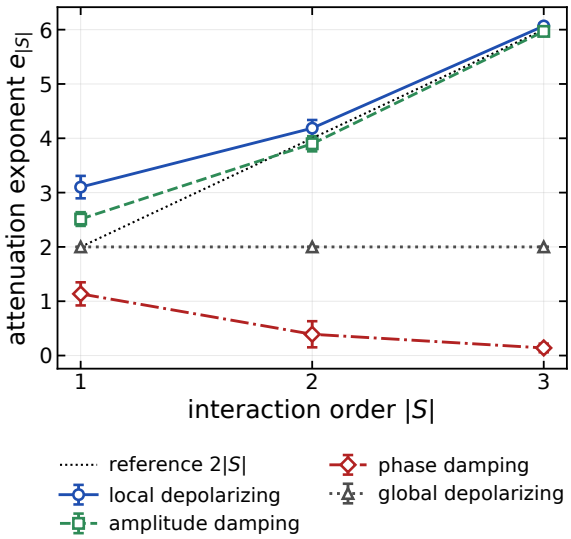


Figure 2: Fitted attenuation exponents by interaction order for four noise channels at m=1.

Table 2: Fitted attenuation exponents by channel and interaction order at m=1. Each exponent is the least-squares slope of $\log(D^{\mathcal{N}}_{|S|}/D_{|S|})$ against $\log q$ through the origin over $p \in \{0.01, 0.03, 0.1\}$, with values reported as mean $\pm$ standard deviation over five seeds. The final column gives the slope of the mean exponent versus interaction order.

Channel	e_1	e_2	e_3	Slope
Local depolarizing	3.10 ± 0.21	4.19 ± 0.15	6.07 ± 0.04	$+1.48$
Amplitude damping (non-unital)	2.51 ± 0.13	3.90 ± 0.14	5.97 ± 0.07	$+1.73$
Pure dephasing	1.14 ± 0.21	0.39 ± 0.24	0.14 ± 0.08	-0.50
Global depolarizing (control)	2.000 ± 0.000	2.000 ± 0.000	2.000 ± 0.000	0.00

Appendix D. On trained models across twelve datasets and five seeds, the simulated noisy circuit reproduces the coefficient map $m_P \mapsto \Lambda(P)\, m_P$ to within 3.3×10^{-12}. Across 14,742 subset evaluations under depolarizing and dephasing noise, neither (b) nor (c) is violated, and R^Q_k decreases in every depolarizing run. The sign condition holds for only 5.2% of subsets; among the remainder, the bound $q^{2|S|} D_S$ fails for 15% under depolarizing noise. Thus, the sign condition is relevant in practice. For re-uploading circuits and noise inside W, the exponents below are measured rather than derived.

The attenuation need not increase with interaction order (Figure 2 and Table 2). Under a global depolarizing channel $\rho \mapsto (1-p)\rho + p\, I/2^n$, the decision function transforms as $f \mapsto (1-p)f + c$. Hence every D_S scales by $(1-p)^2$, independently of $|S|$, and all ratios, residuals, and rankings are invariant. Here, the maximum deviation in R_k is only 5.6×10^{-16}. This control, however, cannot reveal order selectivity. Part (a) of Theorem 3 instead shows that each Pauli string is weighted by its own eigenvalues. Under pure dephasing, $\lambda_Z = 1$, so strings supported on Z are not attenuated, producing the reversal observed in Table 2. Randomized compiling (Wallman & Emerson, 2016) can bring device gate noise into Pauli form, but noise inside the trainable circuit changes M_{eff} itself and is therefore outside the theorem.

Corollary 4 (Anti-conservatism of the residual)**.** *Write $A = \sum_{|S| \le k} D_S$ and $B = \sum_{|S|>k} D_S$, so that $R_k = B/(A+B)$. Under any attenuation $\gamma_{|S|}$ decreasing in $|S|$, every term of B has strictly higher order than every term of A, so $R_k^{\text{noisy}} < R_k^{\text{ideal}}$.*

Theorem 3 alone does not imply the monotonicity required by Corollary 4. The measured exponents in Table 2 increase with order for depolarizing and amplitude-damping noise, supporting the corollary in those cases. Pure dephasing provides the opposite behavior, and the residual bias therefore reverses.

Thus, for hardware noise with increasing attenuation by order, the completeness residual under-reports structure beyond order k, making the model appear more pairwise-additive than it is.

5.1 Bias correction

Model $D_S(p) = q(p)^{e_{|S|}} D_S^{\text{ideal}}$ with an exponent determined by the ansatz and noise channel rather than the data. Measurements at two noise levels then identify the exponent without a noiseless reference,

$$e_{|S|} = \frac{\log\big(D_S(2p)/D_S(p)\big)}{\log\big(q(2p)/q(p)\big)}, \qquad \widehat{D}_S^{\text{ideal}} = D_S(p) \,/\, q(p)^{e_{|S|}}. \tag{3}$$

Two grid evaluations suffice, compared with three for Richardson-extrapolated zero-noise extrapolation (Temme et al., 2017; Li & Benjamin, 2017). Noise scaling can be obtained through randomized compiling or identity insertion, and q through cycle benchmarking (Erhard et al., 2019). Dividing by $q^{2|S|}$ under-corrects when the measured exponent exceeds $2|S|$. Since calibration error propagates with interaction order, a relative error δ_q limits the order at which the correction remains reliable.

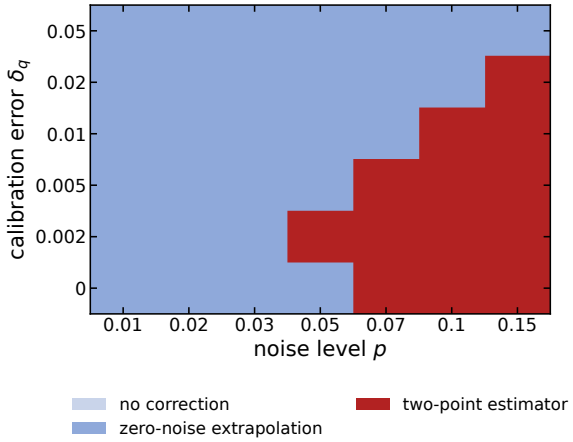


Figure 3: Correction method with the lowest mean absolute R_2 error across noise levels and calibration errors. Zero-noise extrapolation is preferred at lower noise, while the two-point estimator becomes preferable at higher noise when q is well calibrated.

Figure 3 shows the resulting correction regimes. The calibrated devices operate at two-qubit gate error rates of approximately 2.3-2.6 $\times 10^{-3}$, well below the regime where the two-point estimator becomes preferable; we therefore recommend zero-noise extrapolation at these error rates. The two-point estimator is more effective at higher noise. At $p = 0.1$, it gives an R_2 error of 0.0018 ± 0.0011, compared with 0.0027 ± 0.0008 for zero-noise extrapolation and 0.0334 ± 0.0087 without correction. Dividing by $q^{2|S|}$ gives 0.0109 ± 0.0050 (Table 9 in Appendix K). Shot noise biases each D_S upward, so the exact correction in Appendix F is applied before forming any ratio.

6 Experiments

6.1 Setup

We use six-qubit circuits with three trainable layers and a ring IsingZZ entangler: a single-encoding circuit, re-uploading circuits with $m \in \{2, 3\}$, and a two-encoding fidelity kernel with a support-vector read-out. The kernel coefficients are fitted to the noiseless Gram matrix and then evaluated through the device model, corresponding to the setting faced by an auditor of a deployed model. The per-dataset results in Appendix I use eight qubits, the light-cone ablation uses rings of eight and ten qubits, and the hardware experiment uses a six-qubit chain.

We compare against Integrated Hessians (Janizek et al., 2021), the Shapley interaction index (Grabisch & Roubens, 1999), Shapley-Taylor (Sundararajan et al., 2020), Sobol indices (Sobol', 1990), and

Table 3: Comparison of pairwise interaction estimators on twelve datasets and 60 trained models. Agreement is Spearman correlation with the exact D_{ij}; stability is the correlation between noisy and noiseless rankings at three times the device error. Values are mean ± standard deviation over models.

Estimator	Agreement	On device	Stability
Exact grid (target)	1	Yes	0.998 ± 0.002
Integrated Hessians	0.80 ± 0.13	No (needs gradients of f)	
Shapley interaction index	0.72 ± 0.13	Yes	0.997 ± 0.004
Mixed-derivative probe	0.63 ± 0.19	No (needs gradients of f)	
Sampled Sobol, 64 evaluations	0.06 ± 0.23	Yes	0.997 ± 0.006

a naive mixed-derivative probe, all evaluated on the same trained f. Unless stated otherwise, experiments use five seeds and report mean ± standard deviation; the hardware experiment uses three seeds. Twelve datasets cover network intrusion, health, finance, physics, chemistry, and social domains under a fixed pipeline without dataset-specific tuning. Appendix I gives the per-dataset results, and Appendix G gives the full protocol.

Noisy evaluations use a density-matrix model constructed from published `ibm_fez` calibration data with binomial finite-shot sampling. Physical-device measurements are limited to Section 4.1.

6.2 Estimator comparison

Recovery degrades with encoding degree, with different estimators showing different sensitivity. On planted-interaction tasks, Integrated Hessians decreases from 0.81 ± 0.03 to 0.63 ± 0.05 in top-K recovery as the encoding degree increases from one to three, compared with 0.80 ± 0.03 to 0.71 ± 0.02 for our estimator. Full pairwise attribution takes 57 ± 2 ms for our estimator versus $16{,}106 \pm 947$ ms for Integrated Hessians.

For the twelve real datasets, the target is the exact D_{ij}. Table 3 therefore measures agreement with this quantity, device evaluability, and ranking stability under device noise. Sampled Sobol gives an agreement of 0.06 at the 64-evaluation budget used in the comparison. This reflects the sampling budget rather than a limitation of the estimator: under its own product-uniform sampling law, agreement reaches 0.61 ± 0.21 at 729 evaluations and 0.85 ± 0.11 at 4096. The degree bound therefore provides exact identification at a fixed budget rather than making sampling incapable of estimating the same quantity.

The evaluation-based estimators retain their rankings under device noise, with mean Spearman correlation at least 0.997 even at three times the device error. Thus, the effect described by Corollary 4 is primarily a magnitude bias rather than a reordering of interactions. Integrated Hessians and the mixed-derivative probe cannot be evaluated from device measurements because they require gradients of f. The estimators that are closest to the exact variances on clean data therefore require information unavailable from the deployed hardware.

6.3 Residual bias

For the variational circuits, the completeness residual is $R_2 = 0.028 \pm 0.027$, 0.041 ± 0.021, and 0.028 ± 0.019 for m=1, 2, and 3, respectively, with $R_3 \leq 0.002$. The fidelity kernel contains substantially more higher-order structure, with $R_2 = 0.28 \pm 0.09$ and $R_3 = 0.10 \pm 0.05$ across datasets. Across the noise ladder, the contraction follows a power law in q with $R^2 \geq 0.9996$ at every order for channels with a defined Pauli eigenvalue. The fitted exponents increase with order except for dephasing and the global-depolarizing control.

Under the calibrated hardware model, R_2 is biased downward in all 60 single-encoding runs across the twelve datasets, both at the device error and at three times that error. A one-sided sign test over datasets gives $p = 2.4 \times 10^{-4}$. The bias is $2.3 \pm 0.7\%$ of R_2 at device error and $6.8 \pm 1.9\%$ at three times device error. The fidelity kernel, which has greater higher-order mass, shows a larger bias of $3.6 \pm 0.9\%$ and $10.8 \pm 2.5\%$, respectively.

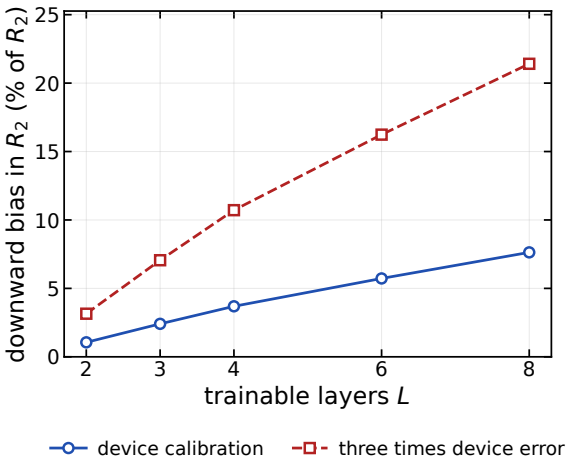


Figure 4: Downward bias in R_2 versus circuit depth at device error and three times device error.

The bias increases with circuit depth (Figure 4), from $1.1 \pm 0.3\%$ at two layers to $2.4 \pm 0.7\%$ at three layers and $7.6 \pm 2.2\%$ at eight layers. At three times the device error, it reaches $21.4 \pm 5.6\%$ at eight layers. The fitted e_2 also increases with depth, from 4.42 to 5.38, while the order-selectivity slope decreases slightly. Thus, most of the depth dependence arises from the increasing higher-order mass rather than stronger order selectivity. The same pattern holds for $m{=}2$: the bias increases from $1.3 \pm 0.3\%$ at two layers to $6.7 \pm 2.0\%$ at eight layers, and from $3.8 \pm 0.7\%$ to $18.6 \pm 5.2\%$ at three times device error (Table 8 in Appendix J).

Limitations. Exact low-order variances scale polynomially with n, whereas exact evaluation of R_k requires an exponential grid and is therefore the main computational cost. Theorem 3 assumes a single encoding layer with noise applied afterward. For re-uploading circuits and per-gate noise, the attenuation exponents are measured rather than derived, and Corollary 4 depends on their monotonicity. This condition holds for depolarizing and amplitude-damping noise but fails for dephasing. At current device error rates, the bias is approximately two percent, with its practical significance increasing with circuit depth as shown in Figure 4. The hardware evidence comes from a single pre-registered six-qubit run on one device; all other noisy results are obtained from a calibrated device model.

7 Conclusion

We presented an interaction attribution framework that accounts for the encoding, architecture, and noise of a quantum model. The finite harmonic structure imposed by the encoding enables exact identification of low-order Sobol variances on a fixed grid, while the entangler graph determines the support of architecture-induced interactions and allows this background to be calibrated independently of the trained model. We further showed that local noise can alter interaction variances through order-dependent attenuation: higher-order interactions are attenuated more strongly under depolarizing and amplitude-damping noise, leading to a downward bias in the completeness residual, whereas this ordering can fail for channels such as pure dephasing. Hardware measurements confirm the predicted interaction reach, and the proposed correction methods provide practical ways to account for noise, with zero-noise extrapolation effective at current device error rates and the two-point estimator applicable at higher noise when the relevant calibration is sufficiently accurate. Together, these results provide a circuit-aware approach to assessing interaction structure in deployed quantum models and to separating learned interactions from effects introduced by the architecture and device.

AI use statement

Generative AI tools were used in this work to assist with grammar and language editing, formatting, figure preparation, and editing of software code. They were also used to help identify potential

mathematical or internal inconsistencies in the manuscript and to provide feedback during the preparation of figures and presentation of results. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of the work, including its methods, results, claims, and conclusions.

Ethics statement

This work uses public benchmark datasets only and involves no human subjects or personally identifying data. Three datasets are network-intrusion benchmarks; the proposed method audits the interaction structure of a deployed classifier rather than performing or facilitating intrusion. Interpretability methods are inherently dual-use, since knowledge of detector feature interactions could potentially be used to evade detection. In this work, attribution requires white-box access to the deployed circuit and its encoding specification.

Reproducibility statement

Code for all experiments, including the commands used to run seeds 0-4, is available in the anonymous repository `https://anonymous.4open.science/r/ORIAN`. Appendix G specifies the models, training settings, preprocessing, and noise model. Appendix I lists the datasets, while Appendices A, B, C, and D provide the complete proofs and their assumptions. Appendix H specifies the hardware experiment, including the predictions registered before the device job was submitted.

References

Armando Angrisani, Antonio A. Mele, Manuel S. Rudolph, M. Cerezo, and Zoë Holmes. Simulating quantum circuits with arbitrary local noise using pauli propagation. *PRX Quantum*, 7:020313, 2026.

Aaron Baughman, Kavitha Yogaraj, Raja Hebbar, Sudeep Ghosh, Rukhsan Ul Haq, and Yoshika Chhabra. Study of feature importance for quantum machine learning models, 2022. URL `https://arxiv.org/abs/2202.11204`.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, November 2019. doi: 10.1088/2058-9565/ab4eb5.

Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld. Better than classical? the subtle art of benchmarking quantum machine learning models, 2024. URL `https://arxiv.org/abs/2403.07059`.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, sept 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00348-9.

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, August 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-63099-6.

R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. *The Journal of Chemical Physics*, 59(8):3873–3878, October 1973. ISSN 0021-9606. doi: 10.1063/1.1680571.

Alexander Erhard, Joel J. Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban A. Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt. Characterizing large-scale quantum computers via cycle benchmarking. *Nature Communications*, 10(1):5347, November 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13068-7.

Enrico Fontana, Manuel S. Rudolph, Ross Duncan, Ivan Rungger, and Cristina Cîrstoiu. Classical simulations of noisy variational quantum circuits. *npj Quantum Information*, 11(1):84, May 2025. ISSN 2056-6387. doi: 10.1038/s41534-024-00955-1.

Elies Gil-Fuster, Casper Gyurik, Adrián Pérez-Salinas, and Vedran Dunjko. On the relation between trainability and dequantization of variational quantum learning models. *ArXiv*, abs/2406.07072, 2024a.

Elies Gil-Fuster, Jonas R. Naujoks, Grégoire Montavon, Thomas Wiegand, Wojciech Samek, and Jens Eisert. Opportunities and limitations of explaining quantum machine learning, 2024b. URL https://arxiv.org/abs/2412.14753.

Francisco Javier Gil Vidal and Dirk Oliver Theis. Input redundancy for parameterized quantum circuits. *Frontiers in Physics*, Volume 8-2020, 2020. ISSN 2296-424X. doi: 10.3389/fphy.2020.00297. URL https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2020.00297.

Kaitlin Gili, Guillermo Alonso, and Maria Schuld. An inductive bias from quantum mechanics: learning order effects with non-commuting measurements. *Quantum Machine Intelligence*, 6(2):67, October 2024. ISSN 2524-4914. doi: 10.1007/s42484-024-00200-0.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, November 1999. ISSN 1432-1270. doi: 10.1007/s001820050125.

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, March 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0980-2.

Raoul Heese, Thore Gerlach, Sascha Mücke, Sabine Müller, Matthias Jakobs, and Nico Piatkowski. Explaining quantum circuits with shapley values: towards explainable quantum machine learning. *Quantum Machine Intelligence*, 7(1):27, February 2025. ISSN 2524-4914. doi: 10.1007/s42484-025-00254-8.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2282952.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007X237892.

IBM Quantum. IBM Quantum processor calibration data: Heron r2 (ibm_fez). https://quantum.cloud.ibm.com/computers?system=ibm_fez, 2026. Median device calibration; accessed 2026.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, and Elham Kashefi. Classically approximating variational quantum machine learning with random fourier features. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ymFhZxw70uz.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4.

Ying Li and Simon C. Benjamin. Efficient variational quantum simulator incorporating active error minimization. *Phys. Rev. X*, 7:021050, Jun 2017. doi: 10.1103/PhysRevX.7.021050. URL https://link.aps.org/doi/10.1103/PhysRevX.7.021050.

Victor Martinez, Armando Angrisani, Ekaterina Pankovets, Omar Fawzi, and Daniel Stilck França. Efficient simulation of parametrized quantum circuits under nonunital noise through pauli backpropagation. *Physical Review Letters*, 134(25), June 2025. ISSN 1079-7114. doi: 10.1103/j1gg-s6zb. URL http://dx.doi.org/10.1103/j1gg-s6zb.

Giuseppe Montalbano and Leonardo Banchi. Quantum adversarial learning for kernel methods. *Quantum Machine Intelligence*, 7(1):15, February 2025. ISSN 2524-4914. doi: 10.1007/s42484-025-00238-8.

Nikita A. Nemkov, Evgeniy O. Kiktenko, and Aleksey K. Fedorov. Fourier expansion in variational quantum algorithms. *Phys. Rev. A*, 108:032406, Sep 2023. doi: 10.1103/PhysRevA.108.032406. URL https://link.aps.org/doi/10.1103/PhysRevA.108.032406.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, February 2020. ISSN 2521-327X. doi: 10.22331/q-2020-02-06-226. URL https://doi.org/10.22331/q-2020-02-06-226.

Lirandë Pira and Chris Ferrie. On the interpretability of quantum neural networks. *Quantum Machine Intelligence*, 6(2):52, August 2024. ISSN 2524-4914. doi: 10.1007/s42484-024-00191-y.

Luke Power and Krishnendu Guha. Feature importance and explainability in quantum machine learning, 2024. URL https://arxiv.org/abs/2405.08917.

Jun Qi, Chao-Han Huck Yang, Pin-Yu Chen, and Min-Hsiu Hsieh. Random-matrix-induced simplicity bias in over-parameterized variational quantum circuits, 2026. URL https://arxiv.org/abs/2601.01877.

Herschel Rabitz and Ömer F. Aliş. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25(2):197–233, june 1999. ISSN 1572-8897. doi: 10.1023/A:1019188517934.

Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *Phys. Rev. Lett.*, 131:100803, Sep 2023. doi: 10.1103/PhysRevLett.131.100803. URL https://link.aps.org/doi/10.1103/PhysRevLett.131.100803.

Maria Schuld. Supervised quantum machine learning models are kernel methods, 2021. URL https://arxiv.org/abs/2101.11020.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Phys. Rev. A*, 103:032430, Mar 2021. doi: 10.1103/PhysRevA.103.032430. URL https://link.aps.org/doi/10.1103/PhysRevA.103.032430.

Thomas Schuster, Chao Yin, Xun Gao, and Norman Y. Yao. A polynomial-time classical algorithm for noisy quantum circuits. *Phys. Rev. X*, 15:041018, Nov 2025. doi: 10.1103/xct1-7kf2. URL https://link.aps.org/doi/10.1103/xct1-7kf2.

Ilya M. Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 2(1):112–118, 1990. Mathematical Modeling & Computational Experiment (Engl.), 1993, 1, 407-414.

Patrick Steinmüller, Tobias Schulz, Ferdinand Graf, and Daniel Herr. explainable ai for quantum machine learning, 2022. URL https://arxiv.org/abs/2211.01441.

Bruno Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2007.04.002.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sundararajan20a.html.

Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. Error mitigation for short-depth quantum circuits. *Phys. Rev. Lett.*, 119:180509, Nov 2017. doi: 10.1103/PhysRevLett.119.180509. URL https://link.aps.org/doi/10.1103/PhysRevLett.119.180509.

Supanut Thanasilp, Samson Wang, M. Cerezo, and Zoë Holmes. Exponential concentration in quantum kernel methods. *Nature Communications*, 15(1):5200, june 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49287-w.

Joel J. Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Phys. Rev. A*, 94:052325, Nov 2016. doi: 10.1103/PhysRevA.94.052325. URL https://link.aps.org/doi/10.1103/PhysRevA.94.052325.

APPENDIX

A PROOF OF THEOREM 1 AND THE EXACT GRAM MATRIX

(A) Per-coordinate degree. With $\theta_i = \pi(x_i{+}1)/2$ and feature i encoded m_i times by Pauli rotations, each encoding gate contributes frequencies $\{-1, 0, +1\}$ to the matrix elements of $\rho(x, \theta)$ in coordinate i. Composing m_i such gates gives frequency support contained in $\{-m_i, \dots, m_i\}$, so $f(x) = \mathrm{Tr}\{\rho(x, \theta)\mathcal{M}\}$ lies in $\bigotimes_i \mathrm{span}\{1, \sin k\theta_i, \cos k\theta_i : k \le m_i\}$ (Schuld et al., 2021; Gil Vidal & Theis, 2020; Nemkov et al., 2023).

(B) Identification. Let $\phi^i_0 = 1$, $\phi^i_{2k-1} = \sin k\theta_i$, $\phi^i_{2k} = \cos k\theta_i$ for $k \le m_i$, and let V^i be the $(2m_i{+}1) \times (2m_i{+}1)$ matrix of ϕ^i_a evaluated at $2m_i{+}1$ distinct nodes. V^i is invertible for any distinct node set; we use $\mathrm{linspace}(-1, 1, 2m_i{+}1)$, which gives condition numbers 2.4, 14.6 and 122 at $m_i = 1, 2, 3$. Evaluating f on the tensor grid and contracting each axis with $(V^i)^{-1}$ returns the coefficient tensor C exactly.

(C) ANOVA from the coefficient tensor. Split each basis function into constant and centered parts, $\phi^i_a = \mu^i_a + \psi^i_a$ with $\mu^i_a = \mathbb{E}[\phi^i_a]$. Expanding the product over coordinates and grouping terms by which coordinates contribute a centered factor is the ANOVA decomposition. Hence $d^S = C$ contracted with the moment vector on every axis outside S, and

$$D_S = \sum_{\alpha,\beta} d^S_\alpha d^S_\beta \prod_{i \in S} G^i[\alpha_i, \beta_i], \tag{4}$$

where G^i is the centered-basis Gram matrix. With $\theta \sim U[0, \pi]$ and $\mu_k = \mathbb{E}[\sin k\theta] = (1 - (-1)^k)/(k\pi)$,

$$G[\cos a, \cos b] = \tfrac{1}{2}\delta_{ab}, \quad G[\sin a, \sin b] = \tfrac{1}{2}\delta_{ab} - \mu_a\mu_b, \quad G[\cos a, \sin b] = \tfrac{1}{2}(\mu_{a+b} - \mu_{a-b}). \tag{5}$$

G is diagonal *only* at degree one, where it equals $\mathrm{diag}(0, 1/2 - 4/\pi^2, 1/2)$. At degree two and above the largest off-diagonal entry is $4/(3\pi) = 0.424$, so a weighted sum of squared coefficients is not D_S beyond degree one. Row and column zero vanish because $\psi_0 \equiv 0$, which enforces the pure-interaction constraint without slicing.

(D) Exact variance. $\mathrm{Var}(f) = \sum_{S \neq \emptyset} D_S$ is computed as $\mathbb{E}[f^2] - \mathbb{E}[f]^2$ from the same tensor using the uncentered second-moment matrix $G + \mu\mu^\top$ per axis, avoiding enumeration of 2^n subsets. All arithmetic is float64; float32 caps reconstruction at $\approx 10^{-8}$.

B FIDELITY-KERNEL DEGREE

$\mathrm{Tr}\{\rho(x)\rho(x')\}$ is linear in $\rho(x)$. After r encodings the matrix elements of $\rho(x)$ carry frequencies up to $\pm r/2$ in each coordinate, so the fidelity's difference set is $\{-r, \dots, r\}$ and the degree is r. The modulus-square does not double the frequency: for $r{=}1$, $R_Y(\theta)|0\rangle$ has frequencies $\pm 1/2$ and the fidelity has degree one, with degree-two coefficients numerically below $2{\times}10^{-12}$. We use $r{=}2$ throughout. The entangler angle must not depend on the data; a ZZ-feature-map coupling $\mathrm{IsingZZ}((\pi - x_p)(\pi - x_q))$ is not a fixed-degree trigonometric polynomial at all, and a degree-three fit leaves residual 0.65.

C A KNOWN DEGREE BOUND IS NECESSARY

Proposition 5 (A known degree bound is necessary)**.** *Fix a coordinate and a budget of N point evaluations. For the class of functions band-limited only by some B with $2B{+}1 > N$, no estimator*

using those N evaluations determines R_k exactly: there exist f and f' agreeing on every queried point and differing in their order-k Sobol variance.

Proof. Let $x_1, \dots, x_N$ be the queried nodes. The space of trigonometric polynomials of degree at most B has dimension $2B+1 > N$, so the evaluation map to $\mathbb{R}^N$ has a nontrivial kernel. Choose $g \neq 0$ in that kernel and set $f' = f + g$. The two agree at every queried point. Because g has a nonzero Fourier coefficient at some frequency, and D_S is a positive-definite quadratic form in the coefficients restricted to the relevant subspace, f and f' can be chosen with different order-k Sobol variance. □

Numerically, certifying a bandwidth-K function on a fixed five-node grid gives reconstruction errors of 0.42, 1.6 and 2.1 at $K = 2, 5, 10$, whereas the genuine degree-two model reconstructs to 10^{-14}. Given query access, by contrast, a band-limited surrogate fits exactly at every degree (Table 4), and what separates it from our approach is the calibrated background rather than the numerators.

Table 4: Query-matched band-limited surrogate against our method at identical query budgets on planted-interaction tasks, n=6. Recovery is the top-K hit rate of the planted pair; entries are mean ± standard deviation over five seeds, each seed averaging three arms and three entangler strengths. The surrogate reproduces the numerators exactly; the difference is the graph-calibrated background.

degree m	queries	surrogate R^2	surrogate+Sobol	Ours	difference
1	729	1.000000	0.767 ± 0.025	0.789 ± 0.025	$+0.022 \pm 0.050$
2	15,625	1.000000	0.622 ± 0.047	0.722 ± 0.056	$+0.100 \pm 0.046$
3	117,649	1.000000	0.667 ± 0.039	0.700 ± 0.030	$+0.033 \pm 0.030$

D Proof of Theorem 3

Placement. After $R_Y(\theta_k)$ the state of qubit k is $\frac{1}{2}(I + \sin\theta_k X + \cos\theta_k Z)$. A Pauli-diagonal channel maps it to $\frac{1}{2}(I + \lambda_X \sin\theta_k X + \lambda_Z \cos\theta_k Z)$, and because the expectation of a Pauli string factorizes over qubits for a product state,

$$f^{\mathcal{N}}(\theta) = \sum_{P \in \{I,X,Z\}^n} \Lambda(P)\, m_P \prod_{k:P_k=X} \sin\theta_k \prod_{k:P_k=Z} \cos\theta_k, \qquad \Lambda(P) = \prod_{k:P_k \neq I} \lambda_{P_k}, \tag{6}$$

where strings containing Y have zero expectation and the noiseless function is the case $\Lambda \equiv 1$. The channel therefore acts as $m_P \mapsto \Lambda(P) m_P$, and part (a) follows by summing squares over $\mathrm{supp}(P) = S$.

ANOVA decomposition. Under the uniform law on $[0, \pi]$, $\mathbb{E}\cos\theta = 0$, $\mathbb{E}\sin\theta = 2/\pi$, $\mathbb{E}[\sin\theta\cos\theta] = 0$, $\mathrm{Var}\cos\theta = 1/2$ and $\mathrm{Var}\sin\theta = 1/2 - 4/\pi^2$. Write $\sin\theta_k = 2/\pi + \tilde{s}_k$ and $\cos\theta_k = \tilde{c}_k$, with $\tilde{s}_k$ and $\tilde{c}_k$ centered and orthogonal. Expanding each product in Equation (6), a term belongs to f_S exactly when its centered factors occupy S. Such a term fixes a *sector* $a \in \{X, Z\}^S$, the Pauli type on S, while each qubit outside S carries I, or carries X and contributes the constant $2/\pi$; strings with a Z outside S have mean zero there and do not contribute. Collecting terms by sector and by the number t of X factors outside S, and writing $M_{a,t}$ for the sum of m_P over strings of that class,

$$D_S^{\mathcal{N}} = \sum_a v_a \lambda_a^2 \Big(\sum_t (2\lambda_X/\pi)^t M_{a,t} \Big)^2, \qquad \bar{D}_S = \sum_a v_a \Big(\sum_t (2/\pi)^t |M_{a,t}| \Big)^2, \tag{7}$$

with $\lambda_a = \prod_{k \in S} \lambda_{a_k}$ and v_a the product over $k \in S$ of $\mathrm{Var}\sin\theta$ or $\mathrm{Var}\cos\theta$ according to a_k. Cross terms between distinct sectors vanish by orthogonality of $\tilde{s}_k$ and $\tilde{c}_k$.

Parts (b) and (c). Since $0 < \lambda_X \leq 1$, $|\sum_t (2\lambda_X/\pi)^t M_{a,t}| \leq \sum_t (2/\pi)^t |M_{a,t}|$, and $\lambda_a \leq \lambda_S$, which gives (b). If for each sector the $M_{a,t}$ share a sign, the left side is also at most $|\sum_t (2/\pi)^t M_{a,t}|$, and comparison with the noiseless case $\Lambda \equiv 1$ gives (c). Without the sign condition a term with $t \geq 1$ may cancel a term with $t = 0$ in the noiseless circuit; the channel shrinks the former more than the latter, and $D_S^{\mathcal{N}} > \lambda_S^2 D_S$ is possible.

Part (d). Let $A = \sum_{1\le|S|\le k} Q_S$ and $B = \sum_{|S|>k} Q_S > 0$. By (a) under an isotropic channel, $A^{\mathcal{N}} \ge q^{2k}A$ and $B^{\mathcal{N}} \le q^{2(k+1)}B$, so $B^{\mathcal{N}}/A^{\mathcal{N}} \le q^2 B/A < B/A$, and $R_k^Q = B/(A+B)$ is increasing in B/A.

Scope and verification. The argument uses a single encoding layer with the channel after it. For $m > 1$ a feature enters several gates separated by trainable unitaries and the products in Equation (6) no longer factor into single-gate terms, so the exponents reported for re-uploading circuits are measured. Amplitude damping is non-unital and adds an identity component, which the argument does not cover. We verified every part on models trained on the twelve datasets of Appendix I at n=6, five seeds each, under depolarizing and dephasing noise at $p \in \{0.01, 0.05, 0.10\}$, over every subset of order at most three with $D_S > 10^{-10}$. The noisy circuit reproduces Equation (6) to within 3.3×10^{-12}. Across 14,742 subset evaluations there is no violation of (b) or (c). The sign condition of (c) holds for 5.2% of subsets; among the others $D_S^{\mathcal{N}} > \lambda_S^2 D_S$ occurs for 15.0-15.5% under depolarizing noise and 5.4-5.5% under dephasing, where $\lambda_S = 1$. The operator-weight residuals R_1^Q and R_2^Q decrease in every depolarizing run, as (d) requires.

E SHIFT TRANSFER BOUND

Let $\delta = \mathrm{TV}(P_X, Q_X)$ and $\|f\|_\infty \le 1$. For any g with $\|g\|_\infty \le 1$, $|\mathbb{E}_P g - \mathbb{E}_Q g| \le 2\delta$, and applying this to g and g^2 with $g = f - \Pi^P_{\le k} f$, $\|g\|_\infty \le 2$, gives $|\operatorname{Var}_P(g) - \operatorname{Var}_Q(g)| \le 4\delta$. Numerically the constant is attained to 3.998. Converting this into a numerical bound on R_k additionally requires a sup-norm bound on $\Pi^P_{\le k} f$ better than $\|f\|_\infty$, which we do not establish, so the bound is not used numerically.

F SHOT-NOISE BIAS IN D_S AND ITS CORRECTION

D_S is a quadratic form in the coefficients, so an N-shot estimate is biased upward: $\mathbb{E}[\hat{c}^2] = c^2 + \operatorname{Var}(\hat{c})$. With per-node shot variance $\mathrm{varF}_{\mathrm{node}} = \sum_q v_q^2(1 - \langle Z_q\rangle^2_{\mathrm{node}})/N$ and $\hat{C} = V^{-1}\hat{f}$, the covariance is $\mathrm{Cov}(\hat{C}) = V^{-1}\,\mathrm{diag}(\mathrm{varF})V^{-\top}$, so for a quadratic form with per-axis kernels K_{ax},

$$\mathrm{bias} \;=\; \mathrm{Tr}(K\ \mathrm{Cov}) \;=\; \sum_{\mathrm{node}} \mathrm{varF}_{\mathrm{node}} \prod_{\mathrm{ax}} \left[V^{-\top} K_{\mathrm{ax}} V^{-1}\right]_{\mathrm{node}_{\mathrm{ax}},\mathrm{node}_{\mathrm{ax}}}, \tag{8}$$

a separable contraction of the variance tensor with per-axis kernel diagonals, deterministic given the model and N. The bias matters for order separation because it is largest in relative terms at high order, where D_S is smallest, and it moves the ratio D_3/D_2 toward one. Table 5 and Figure 5 give the resulting power analysis for the debiased ratio over five trained models.

Table 5: Shot-noise power analysis for the order-separation measurement, 400 resamples per model, debiased estimator, calibrated `ibm_fez` model. A model resolves the effect when its 95% interval on D_3/D_2 excludes one. Entries are mean ± standard deviation over five trained models, whose exact ratios lie between 0.9706 and 0.9916.

shots	mean D_3/D_2	mean 95% interval	$P(\widehat{D_3/D_2} < 1)$	models resolving
2,048	0.984 ± 0.009	$[0.936, 1.032]$	0.75 ± 0.13	$0/5$
8,192	0.984 ± 0.010	$[0.962, 1.008]$	0.87 ± 0.14	$2/5$
32,768	0.984 ± 0.009	$[0.972, 0.996]$	0.95 ± 0.08	$3/5$
131,072	0.983 ± 0.009	$[0.977, 0.989]$	0.99 ± 0.02	$4/5$

G EXPERIMENTAL PROTOCOL

We use six qubits, one per selected feature, with three trainable layers of $R_Y R_Z$ rotations and a ring IsingZZ entangler at angle $\lambda\pi/2$, where $\lambda = 0.5$. The readout is $\sum_q v_q\langle Z_q\rangle + b$. Models are trained with Adam at learning rate 0.05 for 300 full-batch steps. Re-uploading interleaves $m \in \{2, 3\}$ encodings with the trainable layers. The kernel uses a two-encoding angle feature map

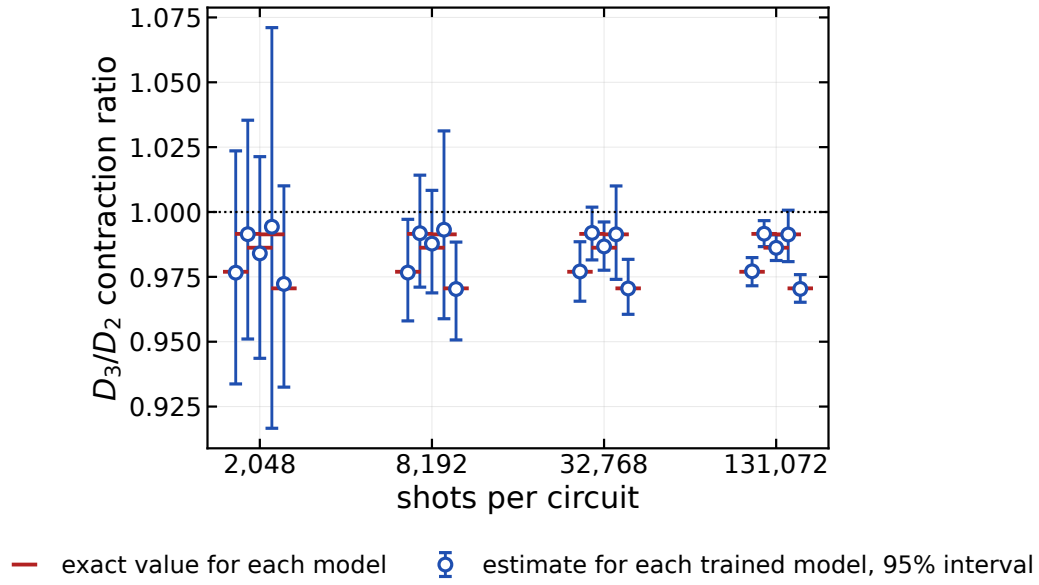


Figure 5: Shot-noise power for the order-separation measurement under the calibrated `ibm_fez` model: debiased D_3/D_2 for each of five trained models with its 95% interval over 400 resamples, and the exact value for the same model. The dotted line marks a ratio of one, the value in the absence of order selectivity.

and a support-vector machine with coefficients fitted on the noiseless Gram matrix and then held fixed. The per-dataset table uses eight qubits; its device-noise results are obtained from a separately trained six-qubit model restricted to the top six features, keeping density-matrix simulation tractable across twelve datasets and five seeds. The light-cone ablation uses rings of eight and ten qubits, while the hardware experiment uses a six-qubit chain.

Datasets are min-max scaled to $[-1, 1]$ using the training split, rank-transformed to uniform marginals, and ranked by mutual information to select the top n features. The per-dataset and device-noise bias experiments use at most 2000 training rows. The certificate, baseline, depth, Sobol-budget, and Theorem 3 verification experiments use the full training split. Controlled experiments on attenuation exponents, estimator degradation, surrogate recovery, null tasks, and shot scaling use synthetic data with planted interactions.

The twelve datasets comprise three network-intrusion benchmarks (NSL-KDD, UNSW-NB15, and CICIDS2017) and nine datasets from health, finance, physics, chemistry, and social domains. We use the same pipeline for all datasets without dataset-specific tuning. All experiments use seeds 0-4. For the per-dataset results, the data split, subsample, feature selection, and initialization are seed-dependent. For the residual-by-class experiments in Table 7, the split is fixed and only the initialization varies with the seed. The fidelity kernel has no trainable initialization, so its results are identical across seeds; its reported spread is therefore across datasets. The hardware experiment uses the three seeds registered before submission.

Noisy evaluation uses a density-matrix model in `default.mixed`, calibrated from published `ibm_fez` data (IBM Quantum, 2026). Thermal relaxation and depolarizing noise are applied after each gate, together with symmetric measurement confusion on each $\langle Z_q \rangle$. Finite-shot measurements are simulated by binomial sampling of each expectation value. Because shot noise biases every D_S upward, particularly at higher orders in relative terms, the exact separable correction is applied before any ratio is reported.

H HARDWARE RUN

The run in Section 4.1 uses the planted function $g(x) = 1.4, x_0 x_4 + 0.5, x_1$ on a six-qubit chain, with labels $y = \mathbf{1}[g > \text{median}(g)]$. The planted pair is therefore at chain distance four. We train

Table 6: All twelve datasets at n=8, mean $\pm$ standard deviation over five seeds.

dataset	accuracy (%)	R_2	conservative	null ratio	device ρ
NSL-KDD	80.0 ± 1.4	0.038 ± 0.012	$94 \pm 7\%$	0.06 ± 0.05	0.998 ± 0.002
UNSW-NB15	92.7 ± 0.2	0.048 ± 0.010	$86 \pm 12\%$	0.04 ± 0.01	0.999 ± 0.002
CICIDS2017	99.8 ± 0.0	0.013 ± 0.007	$83 \pm 7\%$	0.07 ± 0.05	0.999 ± 0.002
Breast-Cancer-W	94.5 ± 2.8	0.029 ± 0.030	$80 \pm 13\%$	0.26 ± 0.12	1.000 ± 0.000
Heart-Statlog	83.5 ± 6.9	0.068 ± 0.021	$87 \pm 13\%$	0.70 ± 0.41	0.999 ± 0.003
Pima	73.9 ± 1.8	0.071 ± 0.027	$68 \pm 12\%$	0.58 ± 0.30	0.999 ± 0.002
German-Credit	70.5 ± 3.2	0.067 ± 0.025	$90 \pm 6\%$	0.23 ± 0.11	0.999 ± 0.002
Default-Credit	81.4 ± 0.3	0.059 ± 0.009	$66 \pm 12\%$	0.27 ± 0.14	1.000 ± 0.000
Adult	83.9 ± 0.4	0.019 ± 0.009	$84 \pm 8\%$	0.09 ± 0.07	0.999 ± 0.002
QSAR-Biodeg	80.1 ± 3.6	0.051 ± 0.019	$64 \pm 11\%$	0.14 ± 0.08	0.997 ± 0.005
MiniBooNE	89.0 ± 0.2	0.027 ± 0.010	$55 \pm 10\%$	0.13 ± 0.05	1.000 ± 0.000
HIGGS	68.0 ± 0.5	0.023 ± 0.007	$51 \pm 13\%$	0.18 ± 0.13	0.998 ± 0.003

circuits at L=2 and L=3 using seeds 0, 2, and 4, while keeping the task, planted weights, training procedure, readout, and physical qubits fixed so that depth is the only difference between the two arms. Seeds were selected in simulation using an accuracy threshold of 0.75 for the L=3 arm. This threshold was not applied to L=2, since that circuit cannot represent the planted interaction and is therefore expected to achieve lower accuracy.

We measure the mixed second difference of the planted pair at $h = 0.8$, averaged over 24 random slices of the remaining four coordinates, using 2048 shots. Each model requires 96 circuits, giving 576 circuits for the complete job. Transpilation targets physical qubits 0 through 5 with optimisation level one and a pinned translation stage. This produces 20 two-qubit gates at L=2 and 30 at L=3, exactly $2(n-1)L$, so the logical chain is mapped without routing overhead. The predictions were registered before submission, and the job consumed 310 of the 600 available quantum-seconds.

I Datasets and per-dataset results

Table 6 lists every dataset with its measured results at n=8 selected features, one qubit per feature, under one fixed pipeline with no dataset-specific tuning. The fields are network intrusion (NSL-KDD, UNSW-NB15, CICIDS2017), health (Breast-Cancer-W, Heart-Statlog, Pima), finance (German-Credit, Default-Credit), social science (Adult), chemistry (QSAR-Biodeg) and physics (MiniBooNE, HIGGS). "Conservative" is the fraction of feature pairs whose grid-computed D_{ij} does not under-report the value under the dataset's own law. The null ratio is the largest interaction of a model trained on shuffled labels, measured beyond the untrained mean, relative to the largest such interaction of the trained model. The device ρ is the Spearman correlation between noiseless and noisy D_{ij} rankings under the calibrated hardware model, computed on a separately trained six-qubit model restricted to the top six features (Appendix G). Reconstruction error stays below 1.1×10^{-14} on every dataset.

The results reveal two distinct effects: the influence of correlated inputs and the effect of device noise. Correlated deployment inputs present a separate challenge from noise: the grid residual is computed under a product law and exceeds the residual under the data's own law in 41 of the 60 runs. Thus, the grid estimate is not conservative in general, with conservatism per pair ranging from 51% to 94% across datasets. In contrast, interaction rankings remain stable under device noise on every dataset, consistent with the distinction between magnitude bias and ranking stability reported in Section 6.

J Residual bias by model class and depth

Table 7 gives the completeness residual by model class. The fidelity kernel carries the most higher-order structure of the four classes, and under the calibrated device model its R_2 bias is the largest, $-3.6 \pm 0.9\%$ at device error and $-10.8 \pm 2.5\%$ at three times it, against $-2.3 \pm 0.7\%$ and $-6.8 \pm 1.9\%$ for the single-encoding circuit.

Table 7: Completeness residual by model class on the twelve datasets. Variational circuits report mean $\pm$ standard deviation over 60 runs (five seeds per dataset); the kernel has no trainable initialization, so its entries are mean $\pm$ standard deviation over the twelve datasets.

model	R_1	R_2	R_3
$m = 1$	0.185 ± 0.109	0.028 ± 0.027	0.0016 ± 0.0023
$m = 2$	0.217 ± 0.096	0.041 ± 0.021	0.0019 ± 0.0015
$m = 3$	0.160 ± 0.066	0.028 ± 0.019	0.0009 ± 0.0006
fidelity kernel	0.527 ± 0.114	0.280 ± 0.090	0.102 ± 0.050

Table 8: Depth scaling of the relative R_2 bias at m=2 under the calibrated device model. Four datasets, five seeds each, twenty complete curves; mean $\pm$ standard deviation. The trend matches the m=1 result of Figure 4.

trainable layers L	bias at $1\times$	bias at $3\times$
2	$-1.28 \pm 0.25\%$	$-3.81 \pm 0.72\%$
3	$-2.08 \pm 0.67\%$	$-6.12 \pm 1.93\%$
4	$-3.20 \pm 0.93\%$	$-9.28 \pm 2.61\%$
6	$-5.14 \pm 1.25\%$	$-14.59 \pm 3.42\%$
8	$-6.68 \pm 2.01\%$	$-18.63 \pm 5.22\%$

K Correction methods

Table 9 compares the correction methods at three noise levels with exact calibration. Dividing by $q^{2|S|}$ under-corrects, because the measured exponent exceeds $2|S|$; the two-point estimator of Equation (3) reads the exponent off the deployed model instead.

L Ablations

L.1 Effect of graph structure and initialization law

Proposition 2 states that the architecture-induced prior depends on the entangler graph rather than the initialization law. Table 10 tests these two factors by varying one while holding the other fixed. Changing the initialization law with the graph fixed keeps the prior within the sampling error of the reference and leaves the score summaries and detection rates close to their matched values. In contrast, changing the graph shifts the prior by more than an order of magnitude, nearly triples the score on null pairs, and more than doubles the false-positive rate. These results show that the graph carries the relevant structural information.

Table 11 isolates the contribution of a background at all. Attribution on raw D_S, with no background subtracted, recovers the planted pair least often. A sampled null and the graph-calibrated prior recover it comparably, and the prior does so without a per-model ensemble. The gain from calibration is real but modest, which is why we describe it as a cost result.

L.2 Effect of entangler structure on light-cone reach

Table 12 varies the entangler while keeping all other settings fixed at two circuit widths. A reach rule is considered valid when no pair with D_{ij} above the numerical floor lies outside the predicted light cone. The three entanglers separate consistently at both widths, indicating that the characterisation is not specific to a particular ring size.

L.3 Operator certificate by dataset

Table 13 gives the per-dataset correlation between the operator certificate and the trained interaction variance, inside the light cone where Proposition 2 is silent. The pre-registered statistic, the median

Table 9: Mean absolute error in R_2 for each correction method at exact calibration ($\delta_q = 0$), mean $\pm$ standard deviation over five seeds. Below $p = 0.05$ zero-noise extrapolation is exact to three decimals, and the two-point estimator overtakes it only at the highest noise level.

method	$p = 0.02$	$p = 0.05$	$p = 0.1$
no correction	0.0069 ± 0.0018	0.0171 ± 0.0044	0.0334 ± 0.0087
division by $q^{2\|S\|}$	0.0022 ± 0.0010	0.0054 ± 0.0025	0.0109 ± 0.0050
zero-noise extrapolation	0.0000 ± 0.0000	0.0003 ± 0.0001	0.0027 ± 0.0008
two-point estimator	0.0001 ± 0.0000	0.0004 ± 0.0002	0.0018 ± 0.0011

Table 10: Initialization-law ablation. Prior discrepancy is the mean relative error of the calibrated prior against a 5000-circuit matched reference, over $m \in \{1, 2, 3\}$. Standardized scores z_S on null and planted pairs are mean $\pm$ standard deviation over five seeds, pooled over m. FPR is the false-positive rate at a threshold fixed on the matched null at m=1; the true-positive rate at that threshold is 1.00 for every law.

law	prior discrepancy	null-pair z_S	planted-pair z_S	FPR
matched	5-11%	638 ± 196	$23{,}720 \pm 10{,}300$	6.2%
Haar	11-12%	742 ± 232	$23{,}600 \pm 6{,}760$	7.7%
anisotropic	16-18%	607 ± 165	$21{,}900 \pm 6{,}510$	4.6%
wrong graph	389-554%	$1{,}754 \pm 769$	$106{,}700 \pm 18{,}000$	15.4%

over datasets of the per-dataset median, is 0.745 against a threshold of 0.70. Eleven datasets clear the threshold individually; UNSW-NB15, with median 0.671 and the largest spread, does not.

L.4 Device-run design for order separation

The order-selectivity effect corresponds to a two percent contraction, and a single trained model resolves it in 41% of draws at 8192 shots. Table 14 varies the number of independently trained models and the shot budget, sampling models at random from five so that the results reflect complete jobs rather than individual draws. Four models at 8192 shots, or three at 32768, resolve the effect in every sampled job. Across the five models, the exact ratio ranges from 0.9706 to 0.9916, with a mean of 0.9833, indicating that the direction is consistent across model instances.

M Extended related work

Interpretability methods for quantum models have so far focused mainly on feature relevance rather than interaction order. Gil-Fuster et al. (2024b) derive main-effect relevance from degree-one Fourier structure and note that their criterion is not expected to extend beyond degree one. This leaves higher-degree settings, including re-uploading circuits and kernels, outside their analysis. Other approaches (Heese et al., 2025; Pira & Ferrie, 2024; Steinmüller et al., 2022; Power & Guha, 2024; Baughman et al., 2022) likewise do not resolve interactions by order. The Fourier structure used here (Schuld et al., 2021; Gil Vidal & Theis, 2020) also provides the basis for classical surrogates of quantum models (Landman et al., 2023; Schreiber et al., 2023). We use such a surrogate as a control rather than as a competing baseline: at matched query budgets, it reproduces our variance numerators exactly. This comparison therefore separates the identification guarantee from the choice of estimator.

The identification result itself is not specific to quantum models. Recovering Sobol indices from a band-limited expansion is a classical problem (Cukier et al., 1973; Rabitz & Aliş, 1999; Sudret, 2008; Sobol', 1990; Hooker, 2007), as is the underlying functional-ANOVA decomposition. Our setting uses the known degree imposed by the circuit encoding to obtain exact identification on a fixed grid. We do not address dequantization (Cerezo et al., 2025; Gil-Fuster et al., 2024a; Gili et al., 2024). Instead, we consider a separate source of structure that is specific to the deployed circuit: the architecture prior. Untrained parameterized circuits exhibit simplicity biases that depend on their architecture (Qi et al., 2026; Thanasilp et al., 2024; Bowles et al., 2024). This motivates calibrating

Table 11: Top-1 recovery of the planted pair with and without a background, on two independent null suites; mean ± standard deviation over five seeds.

attribution	suite 1	suite 2
raw D_S, no background	0.21 ± 0.09	0.16 ± 0.18
sampled null	0.30 ± 0.07	0.34 ± 0.21
graph-calibrated prior	0.29 ± 0.06	0.36 ± 0.22

Table 12: Pairs carrying interaction variance outside the stated rule, out of the measurements listed, on eight- and ten-qubit rings over four depths and five seeds. IsingZZ commutes within a layer and with the readout; IsingXX within a layer only; a CNOT ladder neither.

entangler	width	measurements	$d \le L$	$d \le 2(L-1)$	$d \le 2L$
IsingZZ	$n = 8$	560	20	**0**	0
IsingZZ	$n = 10$	900	75	**0**	0
IsingXX	$n = 8$	560	120	140	**0**
IsingXX	$n = 10$	900	250	225	**0**
CNOT ladder	$n = 8$	560	180	200	60
CNOT ladder	$n = 10$	900	400	375	150

the interaction background from the entangler graph before attributing interactions to the trained model.

Noise introduces a third, distinct consideration. Local noise damps Fourier modes at rates determined by their Hamming weight. Fontana et al. (2025) establish this contraction for Pauli noise with independently parameterized gates, while Angrisani et al. (2026) extend the result to arbitrary local channels and Martinez et al. (2025) to non-unital channels. These results, together with Schuster et al. (2025), use the contraction to truncate high-weight modes and facilitate classical simulation of noisy circuits. Our analysis instead asks how the same attenuation affects interaction variances. The distinction is important because mode-wise contraction does not, in general, imply contraction of an interaction variance in a data-driven circuit: different contributions to the same interaction can cancel in the noiseless model and be attenuated unequally by noise. Section 5 therefore derives the attenuation law at the operator level and identifies the additional condition required for an interaction-order bound.

These structural effects also matter for auditing. An explanation can reflect properties of the model artifact rather than dependencies learned from data, a failure mode documented in both classical and quantum settings (Lapuschkin et al., 2019; Montalbano & Banchi, 2025). The graph-calibrated background in Section 4 addresses this issue by separating architecture-induced interactions from those that remain in the trained model. Together, these lines of work motivate treating identification, architecture, and noise as separate components of interaction attribution rather than as independent sources of estimation error.

N ADDITIONAL RESULTS

Operator certificate. The certificate $\Omega_{i:j}$ is computed from the exact Pauli decomposition of M_{eff}, which reproduces the circuit's own decision function to 5.0×10^{-16}. Out of cone it vanishes: over 135 measurements at distance beyond the cone, Ω stays below 1.8×10^{-29} and D_{ij} below 1.5×10^{-31}, while in-cone medians are 0.897 and 8.2×10^{-3}.

Estimator budgets. Sampled Sobol was given the attribution set, 64 evaluations, in Table 3, against 729 for the exact grid at m=1. On its own product-uniform sampling law its Spearman correlation with the exact D_{ij} is 0.18 ± 0.26 at 64 evaluations, 0.36 ± 0.22 at 256, 0.61 ± 0.21 at 729 and 0.85 ± 0.11 at 4096, over twelve datasets and 60 trained models.

Table 13: Spearman correlation between $\Omega_{i:j}$ and the trained D_{ij} on in-cone pairs, by dataset, over five independently trained models: the median, which enters the pre-registered statistic, and the mean $\pm$ standard deviation.

dataset	median	mean $\pm$ sd	dataset	median	mean $\pm$ sd
Breast-Cancer-W	0.879	0.890 ± 0.038	German-Credit	0.739	0.744 ± 0.052
MiniBooNE	0.843	0.834 ± 0.044	Adult	0.721	0.739 ± 0.052
NSL-KDD	0.843	0.826 ± 0.092	CICIDS2017	0.739	0.724 ± 0.044
HIGGS	0.782	0.781 ± 0.056	Heart-Statlog	0.736	0.716 ± 0.104
QSAR-Biodeg	0.789	0.771 ± 0.156	Pima	0.704	0.707 ± 0.105
Default-Credit	0.750	0.759 ± 0.040	UNSW-NB15	0.671	0.653 ± 0.168

Table 14: Fraction of jobs whose interval on the averaged contraction ratio excludes one, over 200 random draws of M of five trained models ($M = 5$ admits one draw), with the circuit count each design requires.

shots	models	circuits	interval width	jobs resolving
8192	1	625	0.047	0.41
8192	2	1250	0.034	0.57
8192	3	1875	0.029	0.69
8192	4	2500	0.024	**1.00**
32768	1	625	0.025	0.51
32768	2	1250	0.017	0.89
32768	3	1875	0.014	**1.00**

Shot correction and the two-point estimator. At depolarizing $p = 0.05$ and 2048 to 32768 shots, the two-point estimate of D_S with and without the shot correction agrees to within 0.25 percentage points of relative error at every order, so the two corrections do not interfere.