Quantum Geometry Alone Does Not Predict Kernel Utility: Shot- and Noise-Aware Screening for Quantum Kernels

Pablo Rivas

Department of Computing Technology, School of Computer Science and Mathematics, Marist University, 3399 North Road, Poughkeepsie, NY 12601, USA
Pablo.Rivas@Marist.edu

Abstract. Quantum kernels use quantum feature states as similarity functions for classical kernel classifiers, but evaluating many candidate kernels under practical estimation conditions is costly. Geometric difference offers a pre-training measure of how far a quantum kernel is from a classical reference; however, difference in geometry need not imply useful classification behavior. We study shot- and noise-aware screening for four-qubit quantum kernels on five small binary-classification datasets using tuned classical baselines, three quantum feature-map families, exact and finite-shot estimates, simple noise channels, positive-semidefinite (PSD) diagnostics, and downstream support-vector machines. Across 271 completed quantum-kernel evaluations, no candidate exceeded the strongest tuned classical baseline by a balanced-accuracy margin of 0.01. Geometric difference against the best classical reference was often large but was effectively uncorrelated with downstream balanced accuracy (Spearman $\rho = 0.007$), whereas kernel-target alignment was more informative ($\rho = 0.526$). Moreover, 67.3% of raw estimated kernel matrices contained negative eigenvalues. These results show that quantum-kernel geometry alone is insufficient for training decisions and motivate screening that combines geometric, label-aware, and numerical-validity criteria.

Keywords: Quantum kernels · kernel methods · quantum machine learning · geometric difference · kernel-target alignment · noise-aware screening.

1 Introduction

Kernel methods classify data by evaluating pairwise similarities rather than by constructing explicit high-dimensional features. Quantum kernel methods replace a classical similarity with one induced by a circuit-based data encoding: an input x is mapped to a state $|\psi(x)\rangle = U_\phi(x)|0\rangle^{\otimes d}$, and the fidelity kernel is

$$k_Q(x_i, x_j) = |\langle\psi(x_i)|\psi(x_j)\rangle|^2 . \tag{1}$$

The resulting Gram matrix can be used in a classical support-vector machine (SVM). This connection places quantum feature maps within established ker-

nel learning while asking whether circuit-defined similarities can provide useful structure not captured by classical kernels [10,5].

A practical obstacle appears before model training: the candidate quantum kernel space can be expensive to test. For a training set, each kernel requires many pairwise circuit evaluations; finite shots and noisy execution multiply the cost. Pre-screening is therefore appealing. Huang et al. proposed *geometric difference* as a comparison between quantum and classical kernel geometries, and subsequent studies have used or implemented it alongside related diagnostics [7,8,4]. A large score can flag a quantum kernel that differs substantially from a classical reference.

The central issue in this paper is that geometry is not utility. A quantum kernel can differ from a classical kernel while failing to align with labels, failing to improve downstream accuracy, or becoming numerically problematic once estimated with finite shots and noise. These concerns matter because the kernel passed to an SVM is an estimated matrix, not merely an ideal mathematical object. Prior studies show that noise and statistical estimation affect quantum-kernel behavior, including spectral properties and predictions [6,11,9].

We ask: *Does geometric-difference pre-screening remain a reliable training-decision signal when quantum kernels are evaluated under finite-shot and noisy conditions and compared with strong classical baselines?* We answer through a completed four-qubit evidence set over five binary-classification datasets. The study is designed to test screening behavior, not to claim quantum advantage.

Our contributions are threefold:

- We evaluate quantum-kernel screening under exact, finite-shot, and noisy simulation settings using geometric difference, kernel-target alignment (KTA), and numerical-validity diagnostics.
- We compare screened quantum kernels with tuned classical SVM baselines and show that large geometric difference did not identify useful quantum-kernel gains in the completed evidence set.
- We quantify PSD and conditioning issues in estimated kernels, supporting a multi-criterion screening procedure before quantum-kernel training.

The paper is organized as follows. Section 2 gives the minimum kernel and quantum-kernel background and positions this study. Section 3 defines the evidence set and screening design. Section 4 reports the main findings. Section 5 interprets the result and states its limits. Extended definitions, settings, and secondary analyses are assigned to the supplementary appendix.

2 Background and Related Work

Classical and quantum kernels. For inputs $\{x_i\}_{i=1}^{n}$, a kernel produces a Gram matrix K with entries $K_{ij} = k(x_i, x_j)$. A valid kernel matrix is symmetric and positive semidefinite (PSD), allowing it to be interpreted as inner products in a feature space. An SVM trained on a precomputed kernel uses those similarities in a margin-based classifier [1,2]. Quantum kernel methods retain the classical

SVM training stage but estimate similarities from data-encoding quantum circuits. With d qubits, an input is mapped to $|\psi(x)\rangle = U_\phi(x)|0\rangle^{\otimes d}$ and compared by the fidelity kernel in Eq. (1). Quantum feature maps and quantum-assisted kernel classifiers were formalized by Schuld and Killoran and demonstrated for supervised learning by Havlíček et al. [10,5].

Pre-screening candidate kernels. Computing a full kernel matrix is expensive because each candidate feature map requires pairwise kernel evaluations for both training and testing. It is therefore useful to reject weak candidates before a larger run. Geometric difference measures how unlike a quantum Gram matrix is from a classical reference after regularized normalization. We use

$$g(K_Q, K_C) = \left\| K_Q^{1/2} (K_C + \epsilon I)^{-1} K_Q^{1/2} \right\|_\infty^{1/2}, \tag{2}$$

where $\epsilon > 0$ stabilizes inversion. This is a geometry comparison, not an accuracy estimate. Huang et al. introduced geometric comparisons for studying when quantum and classical kernel models may differ [7]. QuASK implements geometric difference and alignment among quantum-kernel evaluation tools [4], and Krunic et al. applied geometric difference to quantum kernels for electronic health-record prediction [8].

A complementary label-aware measure is kernel-target alignment (KTA). For labels $y \in \{-1, +1\}^n$,

$$\mathrm{KTA}(K, y) = \frac{\langle K, yy^\top \rangle_F}{\|K\|_F \, \|yy^\top\|_F}. \tag{3}$$

KTA increases when the similarity structure agrees with the labels; it is a known classical kernel-analysis measure [3]. Unlike geometric difference, it explicitly uses the supervised target and is therefore a natural comparator for a utility-focused screening test.

Shot/noise effects and kernel validity. With S shots, a fidelity-kernel entry is estimated by a measurement frequency, $\widehat{k}_Q = c_{0\ldots0}/S$. Noisy or independently estimated entries can change the measured spectrum. Since an SVM expects a valid kernel, a practical screen should test PSD structure and conditioning in addition to geometry. Studies of noisy quantum kernels, concentration under statistical estimation, and quantum-kernel matrix processing motivate this requirement [6,11,9].

Gap addressed here. Prior work establishes that circuit-induced kernels can be used by classical learners, defines geometry-based comparisons, and studies ways in which estimated quantum kernels can change under sampling or noise. Those results do not by themselves answer the practical screening decision tested here: when a kernel receives a large geometry-difference score, does it subsequently improve classification over a tuned classical reference under the same estimation conditions? Our work examines this decision directly by joining four evidence

Table 1. Experimental protocol retained in the main paper. Full run identifiers and secondary results are assigned to the supplementary appendix.

Component	Values used in the completed evidence set
Data	D1 moons, D2 circles, D3 Gaussian blobs, D4 breast cancer, D5 wine; four selected/scaled features per instance.
Splits	Stratified 70/30 train/test splits; D1–D2 seeds 0–1, D3 seeds 0–2, D4–D5 seeds 0–2.
Classical references	Tuned linear, polynomial, RBF, and Laplacian SVM kernels; best balanced accuracy per split is the utility reference.
Quantum feature maps	Angle-rotation, IQP-style, and one-layer hardware-efficient data encodings where completed.
Estimation conditions	Exact kernels and completed finite-shot rows at 128, 512, and 2048 shots; clean N0 and simulated-noise N2 settings.
Diagnostics/outcome	Geometric difference, KTA, conditioning/effective-rank checks, negative eigenvalues, PSD-clipped variants, and SVM balanced accuracy.
Decision rule	Useful quantum candidate if $\Delta_{\mathrm{BA}} \geq 0.01$ relative to the best tuned classical SVM on the same split.

types in one experiment: a geometry score, a label-aware score, numerical-validity checks, and downstream SVM utility. The paper therefore does not propose a new kernel score or a general negative result about quantum learning. It tests whether geometry alone is enough to justify training in the completed regime.

3 Study Design and Screening Procedure

3.1 Evidence boundary and data processing

We analyze five binary-classification datasets after reduction and scaling to four real-valued features, matching a four-qubit encoding: moons (D1), circles (D2), Gaussian blobs (D3), breast cancer (D4), and wine (D5). D1–D3 are synthetic controls and D4–D5 are tabular real-data probes. All splits are stratified; preprocessing is fit on the training data and then applied to the test data. D1, D2, D4, and D5 contain $n = 100$ instances, with 70 train and 30 test points; D3 contains $n = 250$, with 175 train and 75 test points.

The completed evidence set combines validated D1–D2 pilot rows, complete D3 stress-test rows for seeds 0–2, and the D4–D5 real-data probe. Incomplete rows from an interrupted larger run are excluded. The resulting corpus contains 271 quantum-kernel SVM evaluations and 780 classical SVM evaluations. This is the evidence boundary for every claim in the paper; it is not the full larger grid originally proposed. Table 1 states the retained protocol, and Table 3 reports the outcome scope.

3.2 Baseline and quantum-kernel models

Classical references are tuned SVMs using linear, polynomial, RBF, and Laplacian kernels. Hyperparameter selection is conducted using training data only, and for each dataset/split the strongest tuned classical balanced accuracy defines the reference utility level. This choice is essential to the research question: a screening signal is practically useful only when the retained quantum candidate competes with a credible classical alternative.

Quantum candidates use fixed, non-trainable data encodings so that the experiment measures kernel selection rather than variational circuit training. The completed runs include: (i) an angle-rotation map with features encoded through single-qubit rotations; (ii) an IQP-style map containing phase encodings and fixed two-qubit coupling structure; and (iii) a one-layer hardware-efficient data map with feature rotations and fixed couplings. D1–D2 contain completed rows for the first two maps, D3 includes all three maps, and D4–D5 contain completed IQP-style and hardware-efficient rows. For a circuit-defined feature state, the train Gram matrix and test-to-train matrix are passed to an SVM as precomputed kernels.

3.3 Kernel estimation and numerical repair

Ideal rows are computed without sampling or injected noise. Finite-shot rows estimate an overlap probability from the frequency of the all-zero outcome:

$$\widehat{k}_{Q,S}(x_i, x_j) = \frac{c_{0\ldots0}(x_i, x_j)}{S}, \tag{4}$$

where S is the shot budget. Completed finite-shot conditions include 128, 512, and 2048 shots. The completed noise settings include a clean simulator condition, N0, and a controlled noisy condition, N2, produced by the simulation pipeline's simple noise model. N2 is a stress condition rather than a calibrated model of a particular device.

An exact ideal fidelity kernel is PSD up to numerical tolerance, but an estimated matrix may contain negative eigenvalues. For an eigendecomposition $K = V\Lambda V^\top$, the PSD-clipped version is

$$K_+ = V\,\max(\Lambda, 0)\,V^\top. \tag{5}$$

We retain raw and PSD-clipped rows as distinct analysis conditions where both were completed. Clipping is not presented as a new method: it is a validity repair used to measure whether the downstream utility conclusion depends on correcting an invalid measured matrix.

3.4 Screening metrics, utility definition, and analysis

For each completed candidate, we compute geometric difference against an RBF kernel and against the strongest classical reference, KTA from Eq. (3), condition

number, effective rank/effective-dimension summaries, and negative-eigenvalue counts. The central downstream quantity is the balanced-accuracy margin

$$\Delta_{\mathrm{BA}} = \mathrm{BA}(K_Q) - \max_{K_C} \mathrm{BA}(K_C). \tag{6}$$

A candidate is called useful at threshold δ only when $\Delta_{\mathrm{BA}} \geq \delta$. The main decision threshold is $\delta = 0.01$, representing a positive one-point balanced-accuracy gain; $\delta = 0$ is reported to expose ties.

We evaluate three questions: (Q1) whether any completed quantum kernel improves balanced accuracy against tuned classical kernels; (Q2) whether geometric difference ranks downstream behavior compared with KTA and numerical diagnostics; and (Q3) whether estimated kernels exhibit PSD/conditioning issues. For Q2, we report global same-condition Spearman correlations between screening metrics and quantum SVM balanced accuracy. The analysis is descriptive over the completed evidence set; it is not a universal hypothesis test over all quantum kernels.

3.5 Reproducibility controls and evidence pairing

Every reported quantum-kernel row is paired with a dataset identifier, train/test split seed, feature-map identity, estimation condition, repair policy, screening scores, and downstream SVM outcome. The strongest classical reference is selected on the same split so that the margin in Eq. (6) compares like with like. The evidence merge retains only completed quantum rows with corresponding screening and classifier records; incomplete rows from the interrupted larger computation are not used to support claims.

Three controls prevent optimistic interpretation. First, classical references are tuned independently of the quantum candidates and include nonlinear kernels. Second, geometric difference is computed before inspecting its relationship to downstream outcome; the reported correlation is a diagnosis of screening behavior rather than a rule chosen to maximize performance. Third, raw and PSD-repaired matrices are distinguished so that a repaired matrix is not confused with the measured raw estimate. These controls make the central null outcome meaningful: the lack of a positive quantum-over-classical margin is not a consequence of comparing against a single weak baseline or ignoring invalid kernel matrices.

3.6 Reporting convention and statistical scope

Balanced accuracy is the primary downstream score because all tasks are binary and the real-data probes need not have perfectly balanced labels. Accuracy and macro-F_1 are retained in the result package as secondary checks. Metric-performance associations are summarized with Spearman rank correlation because the screening values, notably geometric difference and condition number, have broad, skewed ranges. The correlations are computed over the completed rows within the reported evidence boundary. They describe this study; they are not inferential claims about all datasets, encodings, or quantum devices.

Table 2. Operational reading of the pre-screening diagnostics. No diagnostic is treated as a substitute for downstream evaluation.

Diagnostic	Question addressed before broader training	Interpretation in this study
g_{best}	Is the quantum kernel geometrically unlike a strong classical reference?	Large values occurred without positive utility; difference alone is insufficient.
KTA	Does the kernel similarity structure agree with class labels?	Strongest observed rank association with balanced accuracy; useful as supporting evidence.
PSD test	Is the estimated Gram matrix valid for a kernel learner?	Raw shot/noise estimates often failed; repair or rejection is required.
Conditioning	Is numerical instability likely to affect fitting?	Reported as a risk flag, not a utility certificate.
Δ_{BA}	Does training the candidate beat the best classical reference?	Final decision quantity; no row met $\delta = 0.01$.

3.7 Operational screening workflow

The diagnostics are read in a fixed order. A candidate first receives a geometry comparison against a strong classical reference, since a candidate that is merely different from a weak baseline is not informative. It is then checked for label agreement through KTA and for matrix validity through PSD and conditioning diagnostics under the same shot/noise condition that produced the kernel. Only after these checks is its downstream SVM margin compared with the tuned classical reference. Table 2 summarizes this interpretation.

This order prevents two errors. A *false keep* occurs when geometric difference is large but the downstream candidate is not useful; this is the central failure mode observed here. A *false reject* would occur if a candidate with low or moderate geometric difference later produced a useful gain. The completed set does not contain a positive-gain candidate at the primary threshold, so it supports diagnosing false keeps but does not estimate how often a useful candidate might be rejected in a broader search. This asymmetry is why we frame the result as screening evidence, not as a universal selection rule.

4 Results

4.1 Quantum geometry did not yield downstream gains

The completed evidence set contains five datasets and 271 quantum-kernel evaluations (Table 3). At the primary usefulness threshold, $\delta = 0.01$, no evaluated quantum kernel improved balanced accuracy over the strongest tuned classical baseline: 0/271 rows were useful. Even with $\delta = 0$, only four rows matched

Table 3. Completed evidence set and best balanced-accuracy outcomes. Values are mean ± standard deviation across available seeds; gap is the best observed quantum-minus-classical margin per dataset.

ID	Dataset	Q rows	Best classical BA	Best quantum BA	Max gap	Useful	n
D1	Moons	44	1.000 ± 0.000	0.950 ± 0.024	-0.033	0/44	100
D2	Circles	44	1.000 ± 0.000	1.000 ± 0.000	0.000	0/44	100
D3	Gaussian blobs	135	0.965 ± 0.061	0.902 ± 0.056	-0.053	0/135	250
D4	Breast cancer	24	0.985 ± 0.026	0.770 ± 0.082	-0.098	0/24	100
D5	Wine	24	1.000 ± 0.000	0.750 ± 0.043	-0.225	0/24	100
	Total	271	–	–	≤ 0.000	0/271	–

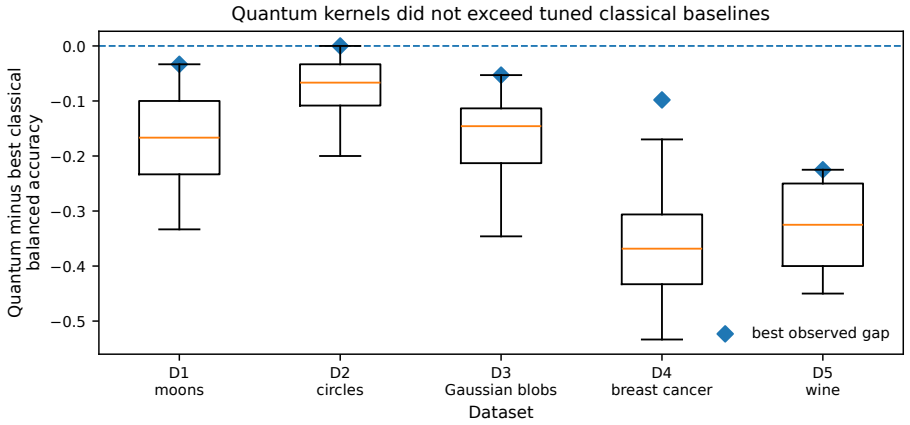


Fig. 1. Quantum-minus-classical balanced-accuracy gaps by dataset. The dashed reference line marks equal performance. No completed quantum-kernel candidate exceeded the strongest tuned classical reference.

rather than exceeded the classical reference. Figure 1 shows the distribution of quantum-minus-classical balanced-accuracy margins. The maximum observed gap was non-positive on every dataset: -0.033 on D1, 0.000 on D2, -0.053 on D3, -0.098 on D4, and -0.225 on D5.

This result does not test whether quantum kernels can be useful in every regime. It establishes the condition required for the paper's screening question: in this completed set, any metric that assigns promise to quantum candidates must be judged against the fact that training those candidates did not provide measured utility over strong classical references.

4.2 Large geometric difference was not a utility signal

Geometric difference against the strongest classical reference was often large: its mean over completed screening rows was 6680.19. Yet its global same-condition association with quantum SVM balanced accuracy was essentially

Table 4. Robustness of the utility and validity conclusions. The utility count compares each quantum-kernel SVM with the strongest tuned classical SVM on the same split. The validity rates refer to completed raw kernel matrices before PSD repair.

Utility margin δ	Compared rows	Useful rows	Useful rate
0.00	271	4	1.5%
0.01	271	0	0.0%
0.03	271	0	0.0%

Raw kernel condition	Rows with negative eigs.
Exact/N0	0.0%
128 shots/N0	100.0%
128 shots/N2	100.0%
512 shots/N0	72.4%
512 shots/N2	86.2%
2048 shots/N0 or N2	100.0%

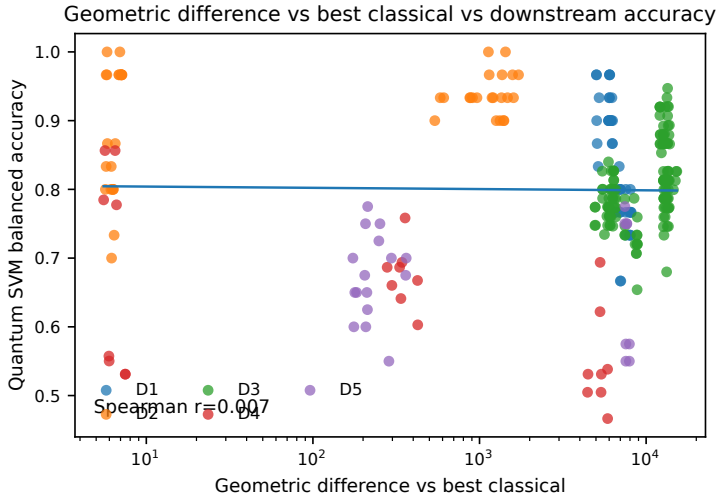
zero (Spearman $\rho = 0.007$, $p = 0.909$). Figure 2a visualizes this central result: high geometric-difference values appear without corresponding gains in balanced accuracy.

In contrast, KTA, which includes label information, had the strongest observed association with quantum balanced accuracy (Spearman $\rho = 0.526$, $p < 10^{-19}$). Figure 2b compares the global rank correlations of the evaluated screening metrics. Condition number and geometric difference against RBF provide weaker positive associations, while effective-rank metrics are negatively associated in this evidence set. These results do not make KTA a sufficient training rule; they show that geometry alone loses the signal most directly tied to supervised utility.

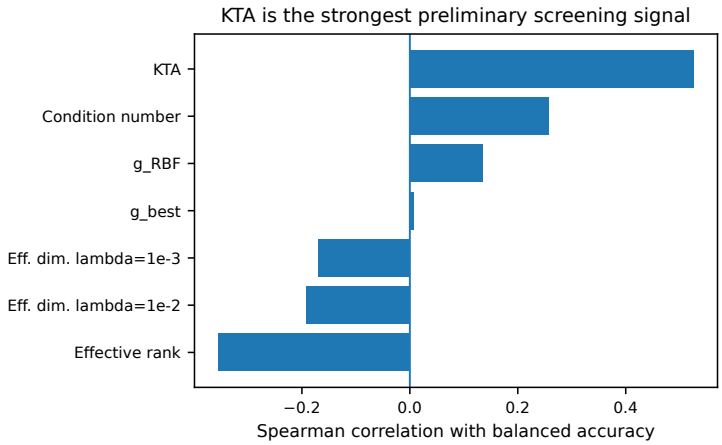
4.3 Label-aware evidence and threshold robustness

The aggregate correlation summary in Fig. 2b is complemented by a direct view of KTA in Fig. 3. Unlike geometric difference, KTA uses the labels and displays a visible positive association with downstream balanced accuracy. This observation does not establish that alignment is a complete selection policy: several rows with comparable KTA still have different downstream performance, and no aligned candidate achieved the primary utility margin. It does establish that the label-aware check preserves relevant information that the geometry-only score did not capture in this corpus.

The outcome is also stable under the utility-margin choice. Table 4 shows that four rows only tie or match the strongest classical reference at $\delta = 0$, while no quantum-kernel row achieves even a 0.01 positive gain. Consequently, the central finding is not driven by an overly demanding improvement threshold.



(a) Geometric difference versus accuracy.



(b) Global metric correlations.

Fig. 2. Screening behavior in the completed evidence set. (a) Large geometric difference against the best classical reference did not imply high downstream balanced accuracy. (b) KTA had the strongest global same-condition correlation with balanced accuracy, whereas geometric difference against the strongest classical reference was nearly uncorrelated.

4.4 Estimated kernels require numerical-validity checks

The third finding is numerical. Exact clean raw kernels showed no negative eigenvalues in the completed rows. Once finite-shot or noisy estimates were used, raw matrices frequently lost PSD structure. Overall, 67.3% of raw estimated kernel rows contained at least one negative eigenvalue; all PSD-clipped groups

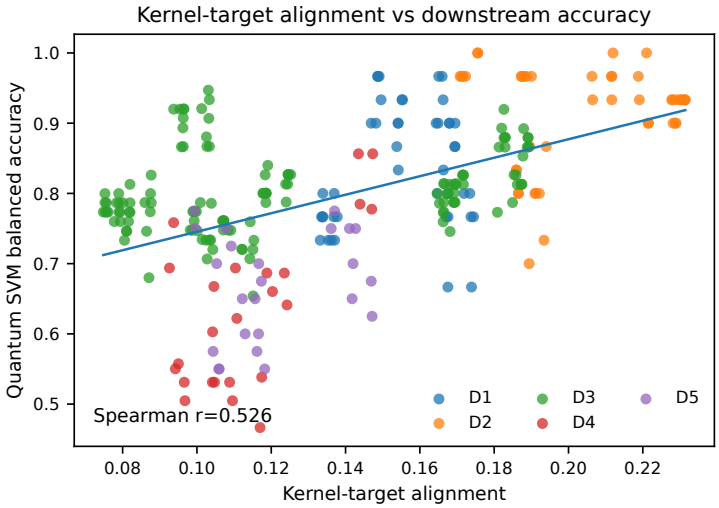


Fig. 3. Kernel-target alignment versus downstream quantum SVM balanced accuracy. KTA is positively associated with downstream performance in the completed evidence set, although this association does not produce a useful candidate against the strongest classical baseline.

had zero negative-eigenvalue rows after repair. Figure 4 shows the negative-eigenvalue distributions for raw matrices by shot/noise condition: raw 128- and 2048-shot groups contain negative eigenvalues in every observed row, while exact clean rows do not. PSD repair makes the downstream matrix valid but does not reverse the utility result: no candidate achieved the primary improvement margin. Thus, numerical repair is necessary for valid training in some settings, but it cannot be treated as evidence that a geometrically distinct kernel is useful.

4.5 Answer to the screening question

Together, the findings answer the three analysis questions. First, the evaluated quantum kernels did not improve on tuned classical baselines at the primary margin. Second, geometric difference alone did not rank useful downstream behavior, while KTA was more aligned with the observed classifier outcome. Third, finite-shot and noisy estimates introduced PSD failures that ideal-kernel screening cannot expose. In this setting, pre-screening should therefore combine geometry comparison with label-aware and numerical-validity checks.

5 Discussion

5.1 What the result means for screening

The title claim is deliberately specific: quantum geometry alone did not predict kernel utility in the completed evidence set. Geometric difference remains

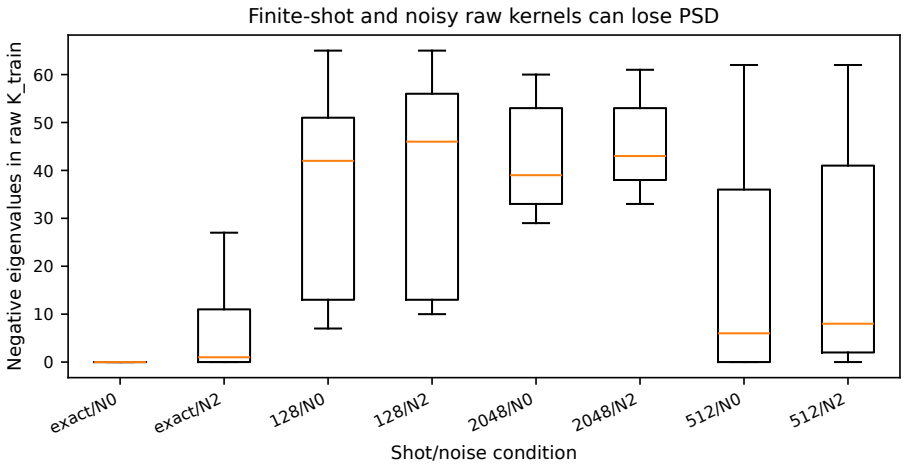


Fig. 4. Negative eigenvalues in raw training kernel matrices by shot/noise condition. Finite-shot and noisy estimation can yield matrices that are not PSD, requiring explicit validation before precomputed-kernel SVM training.

Table 5. Central screening and validity findings. Correlations are global same-condition Spearman correlations with quantum SVM balanced accuracy.

Screening metric	ρ	p
KTA	0.526	$< 10^{-19}$
Effective rank	−0.355	$< 10^{-8}$
Condition number	0.258	$< 10^{-4}$
g_{RBF}	0.134	0.027
g_{best}	0.007	0.909

Validity/outcome fact	Value
Raw rows with negative eigenvalues	67.3%
Exact/N0 raw rows with negatives	0.0%
PSD-clipped rows with negatives	0.0%
Useful rows at $\delta = 0.01$	0/271
Mean g_{best}	6680.19

meaningful as a test of whether a quantum kernel is unlike a chosen classical reference. Its limitation is that it does not include the labels, nor does it establish that a measured matrix is numerically suitable for downstream learning. A candidate can therefore score as geometrically distinct while providing no classification gain. The contrast with KTA, which was more associated with balanced accuracy here, is consistent with this distinction.

The implication is a screening order rather than a new metric. First, compare the candidate against tuned classical references; a weak reference can make a quantum kernel appear more promising than it is. Second, check whether the measured matrix is valid enough for the chosen learning algorithm, especially under shot/noise conditions. Third, use label-aware metrics such as KTA as supporting evidence before spending resources on broader quantum-kernel training. Downstream evaluation remains the final check.

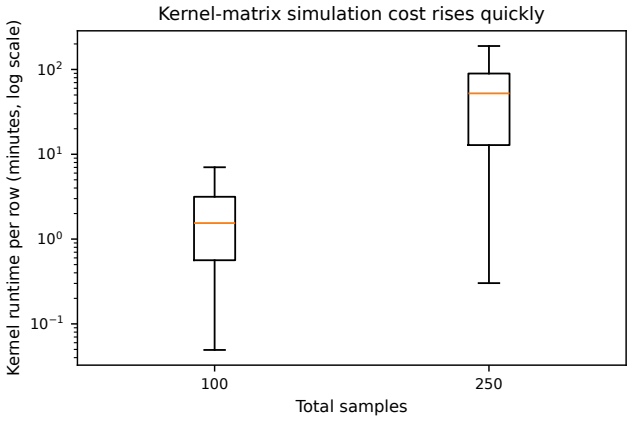


Fig. 5. Per-row quantum-kernel simulation runtime by sample size for completed computations. The observed growth motivates using screening before broader kernel training and reserving the larger planned grid for later work.

5.2 Computational scope and extension to the full grid

The screening problem is not merely statistical: it is computational. A precomputed quantum kernel requires pairwise circuit evaluations for training and test-to-train matrices, and this cost increases rapidly with sample size. Figure 5 reports the observed runtime distribution in the completed corpus. The $n = 250$ D3 block is substantially more costly than the $n = 100$ blocks, which is why the present paper reports completed evidence rather than treating an unfinished larger sweep as a result.

This computational fact strengthens, rather than weakens, the screening question. If producing and evaluating a complete kernel matrix is costly, then a false-positive pre-screening signal spends resources on candidates that do not beat strong classical alternatives. A full extension should avoid redundant quantum computation for post-processed PSD variants, cache raw matrices, test larger sample sizes and higher qubit counts, and include repeated shot realizations and hardware-calibrated noise conditions.

6 Threats to Validity

Evidence-set scope. The reported evidence is a completed subset of a larger intended sweep: five binary datasets, four selected features/qubits, fixed feature-map families, and completed shot/noise settings. It supports the title claim for this regime, but it does not support a claim that all quantum kernels fail or that geometric difference can never be informative. The larger grid is reserved for subsequent work.

Simulation rather than hardware. The study uses exact and noisy simulation and includes no QPU executions. The N2 condition is a controlled stress setting, not a calibrated model of a named device. Therefore, the paper tests screening behavior under controlled degradation and finite estimation rather than reporting hardware-specific accuracy or failure rates.

Model and preprocessing choices. The classical baseline suite is deliberately strong, including tuned nonlinear kernels, because the utility question is whether a quantum candidate earns further training effort. Different dimensionality reduction, encodings, or trainable quantum kernels may change outcomes. Likewise, the present fixed maps separate kernel screening from variational optimization but do not cover all quantum model designs.

Matrix repair and descriptive statistics. PSD clipping changes the matrix supplied to the SVM; it should be interpreted as a validity repair, not a recovery of the raw estimate. Correlation values summarize completed experimental rows and may reflect repeated conditions within datasets. For that reason, we use them as comparative diagnostics, not universal rankings.

The supplementary appendix reports the expanded setting definition, secondary plots, full supporting tables, and further robustness detail. It adds depth while the main argument and limitations remain stated here.

7 Conclusion

This paper reports preliminary evidence that ideal quantum-kernel geometry is not, by itself, a reliable signal of downstream value. In the completed small-data simulations, geometric difference often assigned large values to quantum kernels, yet those kernels did not improve balanced accuracy over tuned classical SVM baselines. At a positive margin of 0.01, none of the 271 quantum-kernel rows was useful relative to the strongest classical reference.

The main lesson is practical. Geometric difference should be treated as a pretraining diagnostic, not as an accuracy proxy. It answers whether two kernels induce different geometries, but it does not say whether the difference helps the label task. In the current evidence set, kernel-target alignment was more tied to balanced accuracy, while PSD and conditioning diagnostics exposed numerical issues in raw finite-shot and noisy kernels.

These findings support a multi-metric screening workflow for quantum-kernel experiments. A candidate quantum kernel should be checked for geometric difference, label alignment, PSD validity, conditioning, and downstream performance under the same shot/noise setting used for training. This does not rule out the usefulness of quantum kernels in other settings, but it does show that ideal-kernel screening can be overly optimistic for small-data classification.

The next step is the full planned grid, including larger sample sizes, more seeds, high-shot settings, and higher-qubit checks. That larger study should also cache raw kernel matrices and derive PSD-repaired variants as post-processing,

rather than recomputing quantum kernels for each PSD policy. The present study gives the core warning and the test workflow needed for that larger run.

Acknowledgments. Part of this work was funded by the National Science Foundation under grant CNS-2136961. The author thanks Dr. Gissella Bejarano Nicho for facilitating access to high-performance computing equipment. The author thanks the Rivas.AI Lab (`https://lab.rivas.ai`) for the support and helpful feedback throughout this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144–152. ACM (1992). https://doi.org/10.1145/130385.130401
2. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**(2), 121–167 (1998). https://doi.org/10.1023/A:1009715923555
3. Cristianini, N., Shawe-Taylor, J., Elisseeff, A.: On kernel-target alignment. In: Advances in Neural Information Processing Systems. pp. 367–374 (2002). https://doi.org/10.7551/mitpress/1120.003.0052
4. Di Marcantonio, F., Incudini, M., Tezza, D., Grossi, M.: Quantum advantage seeker with kernels (quask): A software framework to speed up the research in quantum machine learning. Quantum Machine Intelligence **5**(1), 20 (2023). https://doi.org/10.1007/s42484-023-00107-2
5. Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., Gambetta, J.M.: Supervised learning with quantum-enhanced feature spaces. Nature **567**(7747), 209–212 (2019). https://doi.org/10.1038/s41586-019-0980-2
6. Heyraud, V., Li, Z., Denis, Z., Le Boité, A., Ciuti, C.: Noisy quantum kernel machines. Physical Review A **106**(5), 052421 (2022). https://doi.org/10.1103/PhysRevA.106.052421
7. Huang, H.Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H., McClean, J.R.: Power of data in quantum machine learning. Nature Communications **12**(1), 2631 (2021). https://doi.org/10.1038/s41467-021-22539-9
8. Krunic, Z., Flöther, F.F., Seegan, G., Earnest-Noble, N., Shehab, O.: Quantum kernels for real-world predictions based on electronic health records. IEEE Transactions on Quantum Engineering **3**, 1–11 (2022). https://doi.org/10.1109/TQE.2022.3176806
9. Naveh, A., Fitzgerald, I., Phan, A., Le, A., Kerenidis, I.: Kernel matrix completion for offline quantum-enhanced machine learning. arXiv preprint arXiv:2112.08449 (2021). https://doi.org/10.48550/arXiv.2112.08449
10. Schuld, M., Killoran, N.: Quantum machine learning in feature hilbert spaces. Physical Review Letters **122**(4), 040504 (2019). https://doi.org/10.1103/PhysRevLett.122.040504
11. Thanasilp, S., Wang, S., Cerezo, M., Holmes, Z.: Exponential concentration in quantum kernel methods. Nature Communications **15**(1), 5200 (2024). https://doi.org/10.1038/s41467-024-49287-w

Supplementary Material

To support extension of this study, the supplementary appendix records the dataset blocks, split seeds, feature-map families, shot/noise conditions, PSD policy, outcome threshold, secondary correlations, and runtime behavior. A later full-grid study can therefore extend the same reporting contract while testing larger sample sizes, higher feature/qubit counts, additional shot realizations, and calibrated hardware conditions. This reporting boundary is important: the present title claim refers to completed evaluated conditions, while the supplementary material states exactly what remains to be tested.

The central finding is that geometry alone did not identify downstream utility in the completed evidence set. This appendix adds detail without changing that finding. Section S1 defines classical and quantum kernels, geometric difference, KTA, and numerical-validity diagnostics. Section S2 specifies the completed experiment and the evidence boundary. Section S3 reports supporting results that are informative but not required for the paper's title claim. Section S4 records threshold behavior, compute cost, and limits on interpretation. All results in this appendix refer only to completed rows; incomplete rows from the interrupted larger run are excluded.

S1 Expanded Quantum-Kernel Background

S1.1 Kernels and the quantum-kernel pipeline

A kernel is a pairwise similarity function. For training inputs $x_1, \ldots, x_n$, the Gram matrix is

$$K_{ij} = k(x_i, x_j), \qquad i, j \in \{1, \ldots, n\}. \tag{S1}$$

A valid kernel matrix is symmetric and positive semidefinite (PSD), so that it can be interpreted as inner products in a feature space. In this study, the classifier is a classical SVM supplied with a precomputed Gram matrix. The SVM decision function has the form

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b\right), \tag{S2}$$

where $y_i \in \{-1, +1\}$, and α_i and b are fitted by the SVM. Classical SVMs and kernel methods provide the baseline learning framework used throughout the experiment [1,2].

A quantum feature map replaces a conventional feature map with a circuit. For a four-feature input x, the circuit prepares a four-qubit state

$$|\psi(x)\rangle = U_\phi(x)|0\rangle^{\otimes 4}. \tag{S3}$$

The fidelity quantum kernel compares two inputs through the squared state overlap,

$$k_Q(x_i, x_j) = |\langle\psi(x_i)|\psi(x_j)\rangle|^2. \tag{S4}$$

Thus, the quantum computation supplies similarities, while SVM optimization remains classical. This use of quantum feature states in kernel learning follows established quantum-kernel formulations [10,5].

S1.2 Finite-shot and noisy estimates

An ideal simulator can evaluate Eq. (S4) without sampling error. A shot-based estimate instead repeats an overlap circuit S times and records the all-zero outcome count $c_{0\ldots0}$:

$$\widehat{k}_{Q,S}(x_i, x_j) = \frac{c_{0\ldots0}(x_i, x_j)}{S}. \tag{S5}$$

Every entry in a train or test kernel matrix is then estimated, so entry-wise error can affect the spectrum and the trained classifier. Simple simulated noise channels add a controlled perturbation beyond sampling. Such perturbations do not represent a calibrated hardware device in this study; they test whether an ideal screening score remains informative when the trained matrix is estimated under non-ideal conditions. Prior studies motivate this focus by analyzing noisy quantum kernels and shot-noise-related behavior [6,11].

S1.3 Geometry and label-aware screening

The geometric-difference score used in the experiment compares a candidate quantum Gram matrix K_Q with a classical reference matrix K_C:

$$g(K_Q, K_C) = \sqrt{\left\| K_Q^{1/2}(K_C + \epsilon I)^{-1} K_Q^{1/2} \right\|_\infty}, \tag{S6}$$

where $\epsilon > 0$ stabilizes the inverse and $\|\cdot\|_\infty$ is the spectral norm. This definition is asymmetric: the candidate quantum geometry is evaluated relative to the classical reference. Huang et al. introduced geometric difference to compare possible prediction behavior of quantum and classical kernel models [7]; later work implements or applies it with other quantum-kernel diagnostics [8,4]. A high value denotes geometric separation, not guaranteed classification gain.

Kernel-target alignment (KTA) uses class labels. For binary-label vector $y \in \{-1, +1\}^n$,

$$\mathrm{KTA}(K, y) = \frac{\langle K, yy^\top \rangle_F}{\|K\|_F \, \|yy^\top\|_F}. \tag{S7}$$

KTA increases when pairwise similarity agrees with the label structure. It is a known kernel-analysis quantity, used here as a label-aware comparator rather than as a proposed metric [3,4].

S1.4 Numerical-validity diagnostics

A measured kernel matrix may fail PSD structure even when its ideal kernel is PSD. For a symmetric eigendecomposition $K = V\Lambda V^\top$, PSD clipping forms

$$K_+ = V \max(\Lambda, 0) V^\top. \tag{S8}$$

Clipping permits evaluation with a PSD matrix but changes the measured matrix, so raw and repaired rows are interpreted separately. The analysis records the minimum eigenvalue, the number of negative eigenvalues, and condition number. It also reports two spectrum summaries. If $\lambda_1, \ldots, \lambda_n$ are nonnegative eigenvalues and $p_i = \lambda_i / \sum_j \lambda_j$, effective rank is

$$r_{\mathrm{eff}}(K) = \exp\left(- \sum_i p_i \log p_i \right), \tag{S9}$$

and the regularized effective dimension is

$$d_{\mathrm{eff}}(K; \lambda) = \mathrm{Tr}\big(K(K + \lambda I)^{-1} \big). \tag{S10}$$

These diagnostics characterize the measured matrix supplied to learning; they do not establish utility without downstream evaluation.

S2 Full Experimental Specification

S2.1 Completed evidence boundary

The reported experiment contains five binary-classification datasets reduced and scaled to four real-valued features, corresponding to four-qubit circuit inputs. D1–D3 are synthetic controls; D4–D5 are tabular real-data probes. The retained rows combine completed pilot runs for D1–D2, complete D3 stress-test rows for seeds 0–2, and completed D4–D5 probe rows. Rows from an interrupted incomplete D3 seed are excluded. Table S1 reports the exact data coverage.

Table S1. Completed evidence-set coverage. Each quantum row corresponds to a candidate quantum-kernel SVM evaluation under one completed estimation and repair condition.

ID	Dataset	Type	n	Train/Test	Seeds and quantum maps	Q rows	Classical rows
D1	Moons	Synthetic	100	70/30	0–1; angle, IQP	44	120
D2	Circles	Synthetic	100	70/30	0–1; angle, IQP	44	120
D3	Gaussian blobs	Synthetic	250	175/75	0–2; angle, IQP, HE-1	135	180
D4	Breast cancer	Real-data	100	70/30	0–2; IQP, HE-1	24	180
D5	Wine	Real-data	100	70/30	0–2; IQP, HE-1	24	180
Total completed rows retained for analysis						271	780

The completed evidence set contains 271 quantum-kernel SVM evaluations and 780 tuned classical SVM evaluations. It is not the full originally specified sweep over further sample sizes, more seeds, or higher-qubit checks. All claims in both documents are limited to the completed evidence set.

S2.2 Data handling and baseline reference

All splits are stratified. Preprocessing, including feature scaling and any feature reduction used to obtain four inputs, is fitted on the training split and then applied to the test split. D1, D2, D4, and D5 contain 70 training and 30 test instances; D3 contains 175 training and 75 test instances.

The classical reference set uses tuned SVMs with linear, polynomial, RBF, and Laplacian kernels. Hyperparameter selection uses training data only. For each dataset and seed, the strongest tuned classical balanced accuracy defines the downstream reference. This reference is used both to assess quantum utility and to form g_{best}, the geometric difference against the strongest classical kernel matrix for the same split.

S2.3 Quantum feature maps and estimation conditions

Quantum candidates are fixed data encodings rather than trainable circuits. The angle-rotation map uses feature-controlled single-qubit rotations. The IQP-style map adds phase encoding and fixed two-qubit structure. The one-layer hardware-efficient map combines feature rotations with fixed two-qubit couplings. The completed map availability differs by dataset block, as recorded in Table S1; this reflects completed computation rather than a balanced factorial design.

Table S2 records the retained analysis quantities and non-ideal conditions. The simulation labels N0 and N2 indicate the clean and controlled noisy settings in the completed data. N2 is used as a stress condition; it is not a calibrated model of a named processor.

Table S2. Complete specification of the retained experiment and reported analyses.

Component	Retained value or analysis rule
Features/qubits	Four scaled real-valued features and four-qubit feature-state encoding.
Classical kernels	Linear, polynomial, RBF, and Laplacian SVM references; training-only tuning.
Quantum maps	Angle-rotation, IQP-style, and one-layer hardware-efficient maps where completed.
Kernel estimation	Exact simulation and completed finite-shot estimates at 128, 512, and 2048 shots.
Noise settings	N0 clean condition and N2 controlled simulated-noise stress condition where completed.
Matrix policies	Raw estimated matrices and PSD-clipped variants where both were completed.
Screening scores	g_{RBF}, g_{best}, KTA, condition number, effective rank, effective dimension, and negative-eigenvalue diagnostics.
Downstream outcome	Precomputed-kernel SVM balanced accuracy; accuracy and macro-F_1 retained as secondary outputs.
Utility criterion	$\Delta_{\mathrm{BA}} \geq \delta$ versus strongest classical SVM on the same split; primary $\delta = 0.01$.
Excluded computation	Incomplete larger-run rows; uncompleted larger sample-size and higher-qubit sweeps.

S2.4 Kernel matrices, repair, and decision quantities

For each candidate condition, a square train matrix and a test-to-train matrix are evaluated. The train matrix is screened and supplied to a precomputed-kernel SVM; the paired test matrix is used for prediction. Raw and PSD-clipped rows are separate analysis conditions whenever both were completed.

The primary downstream quantity is the balanced-accuracy margin against the best tuned classical SVM on the same split:

$$\Delta_{\mathrm{BA}} = \mathrm{BA}(K_Q) - \max_{K_C} \mathrm{BA}(K_C). \tag{S11}$$

A row is called useful at margin δ when $\Delta_{\mathrm{BA}} \geq \delta$. The central positive-margin check uses $\delta = 0.01$; $\delta = 0$ and $\delta = 0.03$ are reported for sensitivity. Correlation analyses relate screening metrics to the observed quantum SVM balanced accuracy across completed same-condition rows. These are descriptive comparisons for the reported corpus, not population-level statements about all quantum kernels.

S3 Extended Results

S3.1 Dataset-level screening summaries

Table S3 reports geometric-difference, KTA, condition-number, and negative-eigenvalue summaries by dataset. D1 and D3 show large mean g_{best} values, yet the main paper reports no positive utility margin at $\delta = 0.01$. This supports the interpretation that geometric separation is not enough for a training choice.

Table S3. Screening-metric summary by dataset. Values are means ± standard deviations unless marked as medians or rates.

Dataset	Rows	g_{best}	g_{RBF}	KTA	Cond. median	Cond. q75	Neg.-eig. rows
D1 (moons)	44	6600.47 ± 941.69	825.38 ± 819.79	0.156 ± 0.014	2458.69	4.68×10^4	43.2%
D2 (circles)	44	592.48 ± 637.46	988.73 ± 397.51	0.203 ± 0.021	1626.49	2.46×10^4	43.2%
D3 (Gaussian blobs)	135	$(1.03\times10^4) \pm 3390.40$	2834.82 ± 796.13	0.125 ± 0.039	1.60×10^4	1.11×10^5	46.7%
D4 (breast cancer)	24	1868.75 ± 2465.02	276.22 ± 163.92	0.114 ± 0.017	346.94	639.77	29.2%
D5 (wine)	24	2706.47 ± 3571.10	224.56 ± 55.90	0.121 ± 0.016	201.82	382.76	20.8%

S3.2 Label-aware and numerical metric associations

Figure S1 shows KTA versus balanced accuracy for every completed quantum-kernel SVM row. Figure S2 shows condition number against the same outcome. Table S4 reports all global same-condition correlation summaries used in the analysis. KTA has the highest positive Spearman association in this corpus, whereas g_{best} is effectively uncorrelated with accuracy. The condition-number association is weaker and should be read as a numerical signal, not as a target for selecting models.

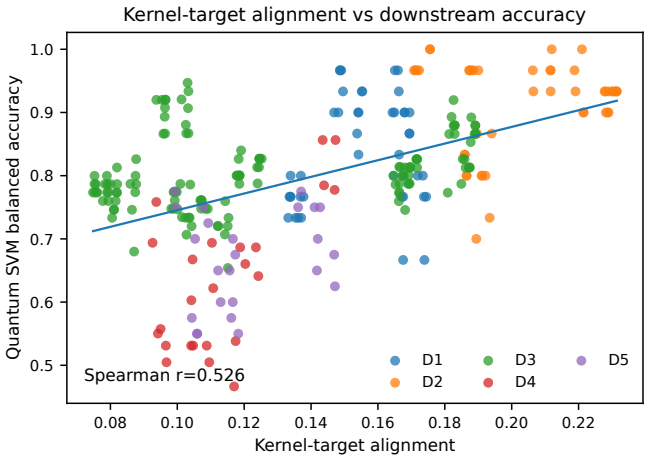


Fig. S1. Kernel-target alignment against quantum SVM balanced accuracy over completed rows. The association is positive in this evidence set (global same-condition Spearman $\rho = 0.526$), but the plot does not imply that KTA alone is a sufficient keep rule.

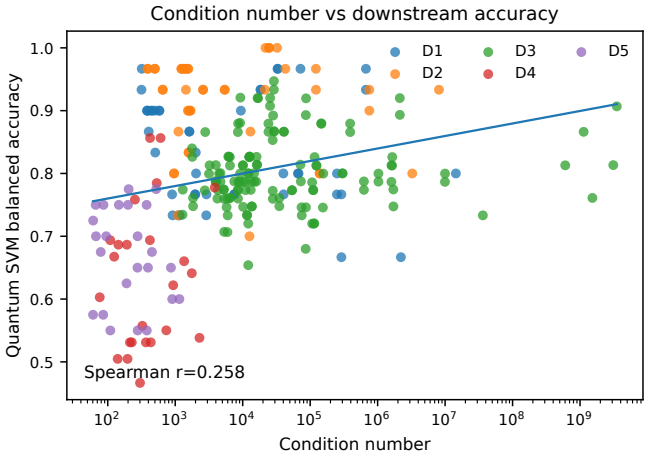


Fig. S2. Condition number against quantum SVM balanced accuracy. Conditioning is included as a measured-matrix validity diagnostic rather than a direct predictor of useful classification.

Table S4. Global same-condition associations with quantum SVM balanced accuracy. All rows contain $n = 271$ completed quantum-kernel evaluations.

Metric	Spearman ρ	p	Kendall τ	Pearson r
KTA	0.526	1.04×10^{-20}	0.357	0.524
Effective rank	−0.355	1.87×10^{-9}	−0.238	−0.312
Condition number	0.258	1.71×10^{-5}	0.170	0.049
Effective dimension, $\lambda = 10^{-2}$	−0.191	0.002	−0.104	−0.052
Effective dimension, $\lambda = 10^{-3}$	−0.169	0.005	−0.084	−0.029
g_{RBF}	0.134	0.027	0.079	0.167
g_{best}	0.007	0.909	0.002	0.055

S3.3 Validity repair and outcome sensitivity

Table S5 reports numerical-validity behavior by shot, noise, and PSD policy, followed by the useful-label sensitivity summary. Exact clean raw rows have no negative eigenvalues, whereas raw shot/noise groups commonly do. PSD clipping removes negative eigenvalues in the repaired matrices; it does not turn any candidate into a positive-margin winner at $\delta = 0.01$.

Table S5. Numerical validity and utility-threshold sensitivity. Panel A summarizes negative eigenvalues and conditioning by completed shot/noise/repair condition. Panel B shows that the utility conclusion is stable across three balanced-accuracy margins.

Panel A: Kernel validity by condition

Shots	Noise	Policy	Rows	Neg.-eig. rows	Mean neg. eigs	Cond. median	Min eig. median
128	N0	PSD-clipped	17	0.0%	0.00	9151.24	-1.22×10^{-15}
128	N0	Raw	17	100.0%	33.35	9235.31	−0.415
128	N2	PSD-clipped	17	0.0%	0.00	2017.50	-1.51×10^{-15}
128	N2	Raw	17	100.0%	35.82	2036.41	−0.453
512	N0	PSD-clipped	17	0.0%	0.00	5316.71	-1.31×10^{-15}
512	N0	Raw	29	72.4%	17.38	2290.45	−0.093
512	N2	PSD-clipped	17	0.0%	0.00	5941.09	-1.51×10^{-15}
512	N2	Raw	29	86.2%	19.31	1355.24	−0.101
2048	N0	PSD-clipped	9	0.0%	0.00	1.48×10^{4}	-1.58×10^{-15}
2048	N0	Raw	9	100.0%	42.56	1.48×10^{4}	−0.108
2048	N2	PSD-clipped	9	0.0%	0.00	2.55×10^{4}	-2.43×10^{-15}
2048	N2	Raw	9	100.0%	45.00	2.54×10^{4}	−0.114
Exact	N0	Raw	29	0.0%	0.00	9398.42	2.32×10^{-4}
Exact	N2	PSD-clipped	17	0.0%	0.00	3.02×10^{5}	-4.43×10^{-16}
Exact	N2	Raw	29	51.7%	9.21	2.46×10^{4}	-2.96×10^{-4}

Panel B: Useful-row sensitivity to the margin threshold

δ	Rows	Useful rows	Useful rate	Maximum margin
0.00	271	4	1.5%	0.000
0.01	271	0	0.0%	0.000
0.03	271	0	0.0%	0.000

S3.4 Accuracy shifts relative to ideal clean evaluation

Figure S3 summarizes balanced-accuracy changes relative to the matching exact/N0 baseline. The spread across conditions shows that shot/noise effects are not uniform across all candidates. This figure supplements the title-level result: once the best classical reference is used, none of these shifts produces a quantum improvement at the primary positive margin.

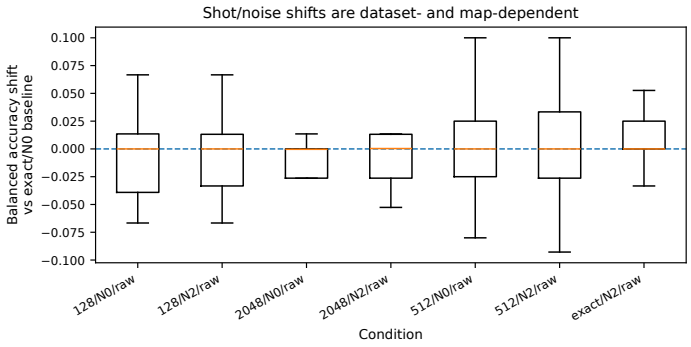


Fig. S3. Balanced-accuracy shift relative to the matching exact/N0 quantum-kernel baseline. Non-ideal estimates can either reduce or alter downstream behavior depending on the candidate condition.

S4 Robustness, Runtime, and Limitations

S4.1 Threshold sensitivity

The useful-label check uses Eq. (S11). At $\delta = 0$, four of 271 completed rows tie or match the strongest classical reference; at the positive margins $\delta = 0.01$ and $\delta = 0.03$, no row is useful (Table S5). Thus, the central conclusion does not depend on requiring an unusually large improvement: even a one-point balanced-accuracy gain is absent in the completed set.

The screening-rule experiment combines thresholds on geometric difference, KTA, and conditioning. Those thresholds are an analysis tool for false keeps and false rejects; they are not proposed as universal operating values. The central evidence remains the downstream margin and the metric associations, which are reported independently of a chosen keep rule.

S4.2 Runtime behavior and scope of the completed set

Kernel matrices require pairwise circuit evaluations for both training and testing, so simulation cost rises rapidly with sample size. Figure S4 shows runtime by

sample size for the completed rows. The $n = 250$ D3 stress-test block is visibly more costly than the $n = 100$ blocks. The larger originally planned grid was stopped after this behavior became clear; only completed seed blocks are included in the reported corpus.

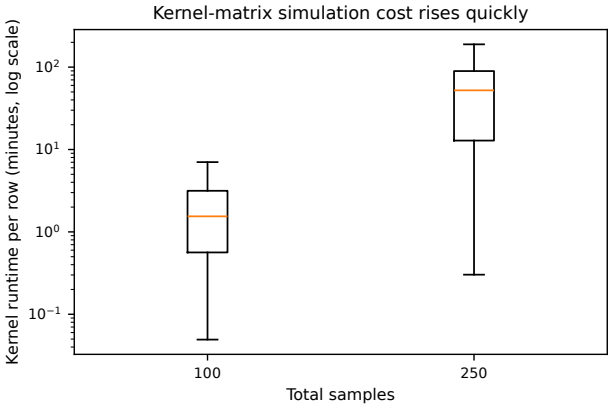


Fig. S4. Per-row quantum-kernel runtime by sample size. The plot records completed computation and motivates separating the reported evidence from larger future sweeps.

A future full-grid study should cache raw quantum matrices and produce PSD-clipped variants by post-processing rather than recomputing the quantum estimates. It should also test larger sample sizes, more shot realizations, higher feature/qubit counts, and additional feature maps.

S4.3 Limits on interpretation

The reported corpus is bounded in several ways. First, it is a completed subset of the larger intended experiment, so it supports a result about the tested regime, not a general statement about quantum kernels. Second, the data are limited to five binary problems and four input features. Third, the feature maps are fixed encodings rather than trainable quantum kernels. Fourth, N2 is a controlled simulation stress condition and is not a device-calibrated noise model. Fifth, the study contains no QPU execution, so it does not claim device-level rates of PSD failure or accuracy shift.

PSD clipping also changes the object supplied to learning. A repaired matrix is useful for testing a valid downstream input, but it is not evidence that the raw estimate was already valid. Finally, the correlation analysis is descriptive over completed rows. It indicates which measured scores align with accuracy in this corpus; it does not prove that one metric dominates in other datasets or circuit families.