METHOD STUDY

The Advantage Island of Quantum Kernels

A measured map for quantum kernel classifiers (QSVM) - where they win, and where they cannot. 996 pre-specified referee arms across a structured factorial grid of label structure, register size and measurement budget, each scored against ten classical variants including an order-matched periodic twin and a tuned RBF kernel SVM, with the study's own screening instruments calibrated against the outcomes. A method study behind the QML RQP.

August 2026

qubit-lab.ch

qubit-lab.ch
contact@qubit-lab.ch

Contents

The study on one page

Theory names two walls for quantum kernel classifiers. If the label structure is simple, classical models reach parity and there is nothing to win. If the register grows, quantum kernel values concentrate exponentially and a device that estimates them from finitely many measurement shots eventually sees noise. Between the walls, if anywhere, lies an island. Each wall has its literature; the map between them had not been measured. We measured it - deliberately on quantum's best case: synthetic datasets whose labels are generated by the quantum kernel itself. The map is a **best-case benchmark envelope** for this feature-map family, not a statement about real-world data.

Throughout this study, "advantage" means a statistically significant predictive advantage against the specified classical referee set under the stated sample and estimation conditions. It does not mean computational speed-up, hardware advantage, or advantage over all possible classical algorithms.

The design: 996 pre-specified referee arms across a structured factorial grid - label structure (prototype density, and a mixing dial toward twin-representable classical structure), register size (4-16 qubits) and kernel estimation (exact statevector; 4,096 shots; 512 shots) - every arm scored by a paired-bootstrap Quantum Advantage Factor against the strongest of ten classical variants, including an order-matched periodic twin, a tuned RBF kernel SVM and a neural network; a win requires the whole 95% confidence interval above 1.0, and a selection-aware robustness bootstrap that recomputes the classical maximum inside every resample confirms 98% of wins. The QML RQP's own screening instruments ran on all 124 study datasets and were scored against the outcomes. The design was pre-specified and locked in writing before the grids ran; every deviation is disclosed in Appendix A.

Four findings survived:

1. **The island is real, and it outlasts the standard per-entry shot heuristic by about two register sizes.** On quantum-native labels the QSVM beats the full referee set in at least 94% of arms at every size up to 12 qubits - at 512 shots as at exact simulation - and erodes only at 14-16 qubits. The common conservative heuristic (kernel entries below 3/sqrt(shots) are unresolvable) predicts failure from about 8 qubits. Ablation experiments locate why it is too pessimistic: sub-threshold entries are individually cheaper to detect than the worst-case variance suggests, the learner aggregates weak signal across hundreds of entries, and the information is redundantly spread across the gram's bulk and its nearest-neighbour tail - with the tail the most noise-robust channel at the largest sizes. This is a quantitative, referee-scored measurement of Thanasilp et al.'s own guidance that the fidelity kernel belongs on data whose embedded states are "not too distant".
2. **The classical wall exists and was located.** Mixing twin-representable low-order structure into the labels kills quantum wins exactly as theory demands: at a pure low-order label the win fraction is 0.000, and wins begin only at 75% quantum-native mixing. The wall was found honestly - the first generator failed its pre-specified validity gate (it produced noise territory, not classical territory) and was revised on the record.
3. **There are two distinct ways to lose to classical, and reports should never conflate them.** On order-1 labels the QSVM learns the task well (macro-F1 near 0.9 throughout) and loses only because classical is perfect - a parity ceiling. On sparse order-3 labels it cannot represent the structure at all. The same distinction cuts the other way: the order-3 twin can represent those labels but does not reliably find them - representability and findability are different properties on both sides of the comparison.
4. **Screening verdicts failed while screening measurements succeeded.** The fast assessment's necessary-condition checklist declared "classical territory" on 83 of 84 quantum-native datasets - including every winner - while the continuous kernel statistics rank the same datasets almost perfectly (AUC up to 0.97, bootstrap 95% interval 0.93-0.99, against shot-feasible wins). The consequence is shipped, not promised: the deep assessment now measures the configured kernel's concentration (bulk and tail, with a tail-based shots-to-resolve estimate) and places each run on this study's benchmark map with calibrated win-fraction bands in place of absolute-threshold verdicts.

Scope, stated once and meant everywhere: kernel methods only (a dedicated follow-up study covers the variational classifier); synthetic quantum-native labels; exact or binomially shot-sampled simulation, no hardware runs; a best-case envelope by construction - our own real-data benchmarks land in classical territory, and detecting exactly that, cheaply and before training, is the practical point.

1. Management summary

Written for decision-makers; the measured basis for every sentence follows in sections 2-8. The three questions below are the questions this study was commissioned to answer.

Should a financial institution invest in quantum machine learning today - and if so, where? We answered this the only way we trust: by measurement. Across 996 pre-specified head-to-head comparisons, quantum kernel classifiers competed against a panel of strong classical methods on data deliberately constructed to give quantum its best possible chance. The result is a map, not a slogan - and the practical deliverable is not a promise of quantum advantage but a measurable search procedure for where further investment is justified, and, more often, where it is not.

Q1 - Where can quantum kernels win at all? In a genuine, bounded region - and finding it is now a search problem, not a leap of faith. There is a region where quantum kernel methods beat strong classical baselines, and it is wider than the skeptics' rule of thumb suggests: in our measurements the method stays viable well beyond the point where textbook arguments predict failure, even under realistic measurement budgets. The region has borders, and that is good news for a decision-maker: borders make the search systematic. Data that lands outside gets a fast, inexpensive "no" - a result that costs days, not a failed pilot that costs a year - and every dataset screened sharpens the picture of where the next serious attempt belongs. The window's upper edge is set by today's measurement budgets, not by principle: as hardware improves, it moves. Measured comprehensively for the kernel method (QSVM); the variational classifier is the subject of a dedicated follow-up study.

Q2 - Can we trust a screening verdict? Only as far as the instruments and budgets behind it - and we quantified exactly that for our own screening instruments: the pre-run checklist would have vetoed almost every dataset that then went on to win, while the underlying measurements rank datasets nearly perfectly - so the product now reports calibrated measurements instead of hard verdicts. A pilot showed the same lesson for training budgets: a quantum "no-go" flipped to statistical parity within a day once the run was configured fairly; the dedicated study follows. The actionable rule for managers: never accept a quantum no-go without checking that the run was fair - and fairness is checkable.

Q3 - What can we order before spending? Suitability, as a measurement. Concretely, a pre-run assessment of a dataset now delivers: (i) its position on the measured advantage region - structure on one axis, problem size on the other, with the boundaries drawn from the benchmark comparisons; (ii) the kernel diagnostics behind that placement, including a concentration reading of the data's own kernel and an estimate of the measurement budget a real device would need to resolve it; (iii) a calibrated confidence band instead of a yes/no - "datasets in this region won the referee comparison in X% of benchmark arms" - with the calibration disclosed; (iv) a method recommendation, including the configuration a fair variational run requires; and (v) a fair-run guarantee - any result obtained under a limiting budget is flagged as such rather than silently reported as a quantum failure. All of it before training, at assessment cost.

One more thing, because it is the standard we believe all quantum claims should meet: in the course of this study we found and published that our own screening instruments were miscalibrated, and we overturned one of our own preliminary conclusions through a pre-specified verification protocol before release. Evidence, not promises - applied first to ourselves.

2. Question and frame

2.1 The question a practitioner actually asks

When is quantum machine learning with kernel methods reasonable? Is there a lower limit - data so simple that a laptop wins - and an upper limit - problems so large the quantum method stops working? And what do the configuration knobs do? The literature offers each wall separately; practice needs the map between them.

2.2 Two walls, one island

The lower wall is classical competitiveness. If a label's structure is representable by accessible classical models - in the spectral picture, if its joint frequency spectrum is sparse or low-order - then classical learners reach parity and no quantum advantage is available in principle [4, 7]. Our referee therefore always contains an order-matched periodic twin: a classical model built on exactly the trigonometric feature classes that a quantum feature map spans at low order.

The upper wall is exponential concentration. Thanasilp, Wang, Cerezo and Holmes proved that under broad conditions the values of quantum kernels concentrate exponentially in the number of qubits toward a fixed value, so that a polynomial number of measurement shots yields a Gram-matrix estimate that is statistically independent of the data, and a trivial model [1]. Rigorous shot-count analyses for fidelity-kernel training reach compatible conclusions [8, 9], and practical shot-estimation heuristics have been proposed [6]. Two properties of these results matter for what follows: the concentration statements are formulated over the ensemble of input pairs - typical-pair statements that leave a measure-small set of atypical pairs unbounded - and the authors themselves direct the fidelity kernel toward data whose embedded states are "not too distant" [1]. Clustered data makes the atypical set non-negligible. This study measures what that means end to end.

2.3 The best-case construction, and why it is honest

Every benchmark dataset in this study carries labels generated by the quantum kernel itself: points are labeled by their nearest class prototype under the exact fidelity kernel the classifier will use (Section 3.1). This is deliberately the best possible case for the quantum method - the problem-inspired embedding of [1] taken to its logical extreme - and it is what makes the map a best-case benchmark envelope for this feature-map family. Inside the winning region, nothing is implied about real-world data; outside the measured region, even this deliberately favourable construction failed to produce an advantage under the tested configuration - a strong empirical signal, though not a universal impossibility claim. Whether a given real dataset carries quantum-reachable structure is a separate, measurable question (Section 7); on our own real-data benchmarks to date the answer has been classical territory.

3. Design

3.1 Generators

Quantum-native family (Phase 1). For a register of n qubits, 2,500 candidate rows are drawn uniformly from [-pi, pi]^n and embedded by the configured feature map (zz-like: RY angle embedding plus CNOT-RZ-CNOT pair terms, two repeats, no input scaling). Each class is represented by k prototypes (k in {1, 2, 4, 6, 8, 12}) drawn from the same distribution; a candidate's label is the class of its nearest prototype by state fidelity, and the 1,000 highest-margin candidates are kept, class-balanced. The label boundary is therefore native to the kernel's own geometry, with k controlling the density of the boundary's joint spectrum.

Mixture family (Phase 2a). A second margin is built from a sparse low-order trigonometric polynomial - three terms, each a product of order_k in {1, 3} single-feature sin/cos factors with on-grid integer frequencies in {1, 2} - and standardized. The label margin is the convex mixture (1 - alpha) x low-order + alpha x prototype margin, alpha in {0, 0.25, 0.5, 0.75, 1}. At alpha = 0 the labels are inside the order-3 twin's representable class by construction; at alpha = 1 the family coincides statistically with the quantum-native family. A pre-specified validity gate required the best referee variant to reach macro-F1 >= 0.85 on the pure low-order labels before any grid run; the first (dense) surrogate failed the gate and was revised to the sparse form on the record (Appendix A).

3.2 The referee

Every arm trains the QSVM (SVC on the precomputed fidelity kernel, validation-tuned decision threshold) and ten classical variants: logistic regression, random forest, gradient boosting and the order-matched periodic twin, each in strict parity and best-reference mode; an **RBF kernel SVM** with (C, gamma) selected on validation over a 3 x 3 grid - the closest classical analogue to a smooth similarity geometry, and the baseline any skeptical reader would demand for prototype-labeled data; and a neural network (two-layer MLP with early stopping). Every variant gets its own validation-tuned decision threshold. The comparison metric is macro-F1 on the held-out test split; the **Quantum Advantage Factor (QAF)** divides the quantum model's macro-F1 by the strongest classical variant's on the same rows, with a paired-bootstrap 95% confidence interval. The verdict rule is absolute: a win requires the entire interval above 1.0; an interval containing 1.0 is statistical parity; an interval below 1.0 is a classical win.

Because selecting the strongest classical variant on the full test set is itself a data-dependent step, every arm additionally carries a **selection-aware robustness check**: a bootstrap in which the classical maximum is recomputed inside every resample, so the claim becomes "quantum beats the referee panel" rather than "quantum beats the variant that happened to be strongest". Both win indicators are reported. Win fractions per cell are counts of whole-interval wins over the cell's six replicates (two generator seeds x three split seeds); the 996 arms form a structured factorial grid, not 996 independent experiments, and are never used as statistical weight.

3.3 Axes and arms

Phase 1: 6 prototype densities x 7 register sizes (4-16 qubits) x 3 estimation levels x 6 replicates = 756 arms. Phase 2a: 2 surrogate orders x 5 mixing levels x 2 sizes (8, 12) x 2 estimation levels x 6 replicates = 240 arms. Fixed throughout: 1,000 balanced rows, 60/20/20 stratified split, identical feature map and training path as the shipping product.

Shot model. Exact arms use the analytic statevector kernel. Shot-estimated arms replace every kernel entry by a binomial overlap-test estimate K-hat = Binomial(S, K)/S at S = 4,096 or 512, symmetrize the train Gram, pin its diagonal, and project it to positive semidefinite by eigenvalue clipping (projection mass recorded; validation and test blocks receive per-entry noise without projection). This models measurement statistics only - it is not a hardware noise model.

3.4 Instrument calibration

All 124 unique datasets (84 quantum-native, 40 mixture) additionally ran the QML RQP's fast assessment (suitability rating, joint-spectrum probe, quantum-territory checklist) and deep assessment (kernel probes, kernel-target alignment, concentration diagnostics, proxy advantage factor). Each candidate diagnostic was scored by ROC AUC against island membership - a cell's win fraction >= 2/3 at the stated estimation level - with axis orientations fixed before any AUC was computed.

3.5 Pre-specification

The design, hypotheses H1-H5 and the analysis plan were written down and locked before any grid run; amendments were recorded before their respective runs. Skeptical readers will find the mechanism's exact scope and limits in Appendix A. All numbers in this document are re-derived from the committed raw results by committed analysis scripts; none are typed in by hand.

4. The map

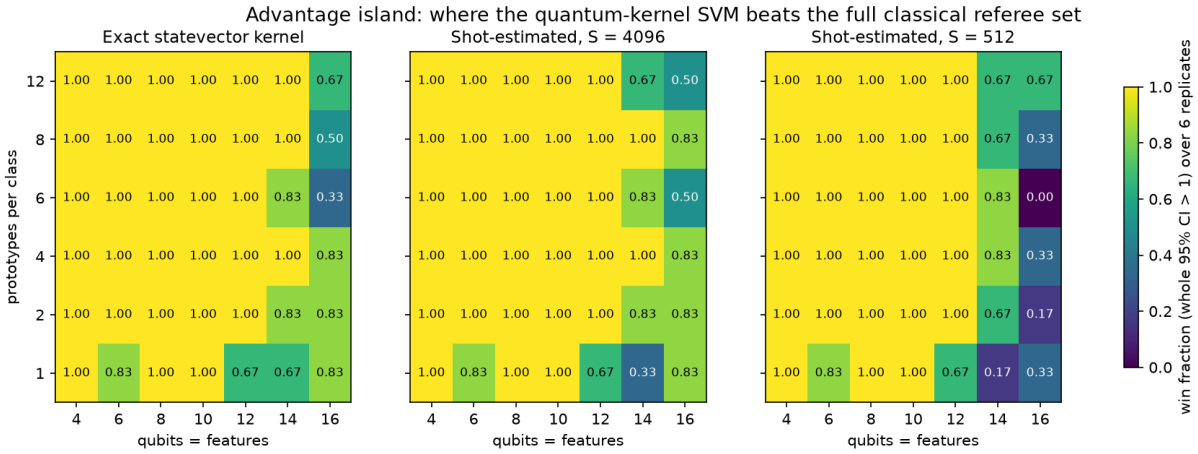

Figure 1 - Win fraction (whole 95% CI above 1.0, over 6 replicates) across prototype density and register size, at the three kernel-estimation levels. The map is a continent through 12 qubits at every prototype density, eroding at 14-16 qubits - hardest at 512 shots, where the 16-qubit column falls to 0.00-0.67.

At exact simulation the QSVM wins at least 94% of arms at every size up to 12 qubits, at every prototype density including a single prototype per class. Median QAF in the strongest cells reaches 1.47 (seed spread 1.28-1.64); even at one prototype the median twin gap is +0.15 macro-F1, rising monotonically to +0.29 at twelve prototypes. The registered

lower-wall hypothesis for this family - that one or two prototypes would be twin-representable - is therefore **refuted**: a nearest-prototype boundary under an entangling feature map is beyond the order-3 twin at every density we generated. The prototype dial moves the margin of the win, not its existence; genuine classical territory required the mixture family of Section 6.

Erosion begins at 14 qubits and is budget-ordered exactly as specified in advance: at 16 qubits the exact-kernel win fraction is 0.67, the 4,096-shot fraction 0.72, the 512-shot fraction 0.31 (with one cell at 0 of 6). A logistic model of the win indicator, with uncertainty from a cluster bootstrap over datasets (accounting for the arms' shared generators and seeds), puts the register-size coefficient at -0.59 per qubit (95% interval -1.03 to -0.40) and the 512-shot penalty at -1.07 (-1.70 to -0.65), while 4,096 shots is statistically indistinguishable from exact (-0.06, interval spanning zero): the paired QAF difference between 4,096-shot and exact arms has median 0.000 (10th-90th percentile -0.032 to +0.017) against -0.008 (-0.081 to +0.020) at 512 shots. The prototype-density coefficient is NOT robust under clustering (interval -0.12 to +0.82) - consistent with the refuted lower-wall hypothesis: within the quantum-native family, density moves margins, not verdicts.

The map is referee-robust. The results above were re-measured with the referee expanded by the tuned RBF kernel SVM and the neural network: not a single one of the 756 arms changed its verdict, and every win fraction in Figure 1 is unchanged to three decimals. The RBF-SVM becomes the strongest classical variant in 12% of arms but never closes the gap: its median macro-F1 falls from 0.87 at 4 qubits to 0.38 at 16 - a smooth stationary kernel on the raw angles cannot track the oscillatory fidelity geometry as the register grows (the neural network sits near chance from 6 qubits). The selection-aware robustness bootstrap agrees with the primary verdicts on 98% of arms (675 primary wins, 667 robust wins, 664 in common).

5. The heuristic wall arrives too early

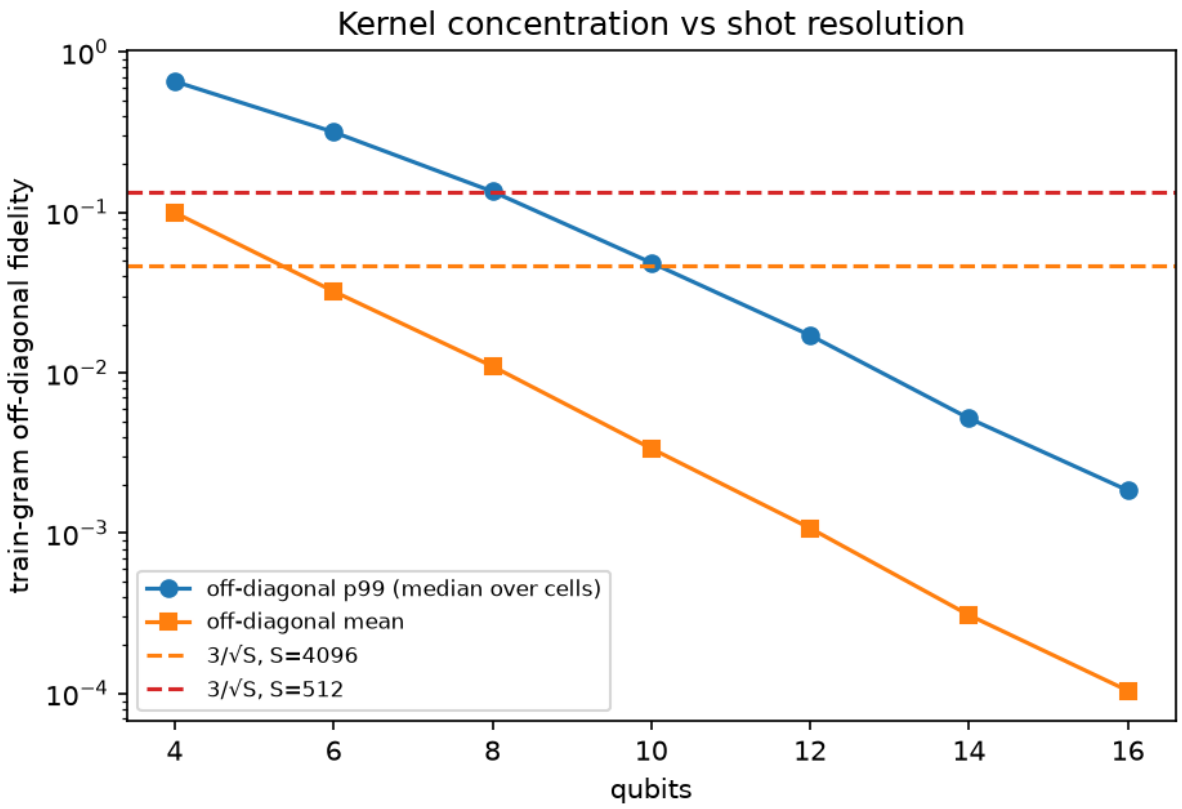

Figure 2 - Median train-Gram off-diagonal statistics against register size (log scale), with the per-entry shot-resolution heuristic 3/sqrt(S). The bulk collapses exponentially exactly as concentration theory predicts and crosses the 512-shot heuristic line near 8 qubits - yet 512-shot arms keep winning through 12 qubits. The wall the referee actually records arrives about two register sizes later than the heuristic predicts.

The bulk of the Gram matrix behaves precisely as [1] predicts: the median off-diagonal kernel value falls from 0.100 at 4 qubits to 1.0 x 10^-4 at 16, and its 99th percentile from 0.66 to 0.0018 - exponential concentration on schedule. A widely used practical heuristic then reasons per entry: an estimate with S shots carries noise of order 1/sqrt(S), so entries below about 3/sqrt(S) are unresolvable, and the method should fail once typical entries sink below that line - near 8 qubits at 512 shots, near 10 at 4,096. The referee disagrees: 512-shot arms win 100% of cells through 12 qubits.

Why the heuristic is too pessimistic - three measured reasons. First, its noise scale is the worst case. The binomial overlap estimator K-hat = Binomial(S, K)/S has variance K(1 - K)/S, not 1/S: small entries are estimated with proportionally small absolute noise, and detecting a small entry against zero requires only that S x K reaches a handful of expected successes - a far weaker condition than the heuristic's absolute threshold. Second, the learner aggregates: a support-vector decision sums signal across hundreds of kernel columns, so per-entry resolvability is not the relevant unit of analysis - the same distinction the theoretical literature draws between estimating individual entries and preserving the information the model needs [1, 8, 9]. Third, the information is redundant, as the ablations below show.

A causal probe of the mechanism. We had conjectured that classification "rides the tail" of large nearest-neighbour fidelities that prototype-clustered data retains (at 16 qubits only ~1% of exact off-diagonal entries exceed 1/512; the 512-shot estimated test kernel is 96% exact zeros - identity-like in the bulk, in the sense of [1]'s trivial-model propositions - yet still correlates 0.85 with the exact kernel). The pre-specified ablation refuted the strong form of that conjecture: on exact kernels, keeping only the top 1% of off-diagonal entries retains most performance (macro-F1 0.86/0.81/0.68 at 8/12/16 qubits against baselines of 0.92/0.88/0.74) - but so does destroying the top 1% and keeping only the bulk (0.83/0.85/0.77). In exact arithmetic the information is redundantly spread across both channels. The tail's special role appears in the shot regime: repeating the ablation on 512-shot estimated kernels, performance degrades gracefully rather than collapsing at the heuristic line, and at 16 qubits the tail-only kernel (0.63) tracks the full estimated kernel (0.66) more closely than the bulk-only kernel does (0.60). A control shows the positive-semidefinite projection applied to noisy grams is not doing hidden work: raw and projected grams give identical results to four decimals, with negative-eigenvalue mass near zero.

None of this contradicts the theorems. The concentration statements of [1] are formulated over the ensemble of input pairs, and their trivial-model conclusions take that concentration as a premise; clustered data violates the premise on the atypical set while satisfying it on the bulk. Indeed the authors' own discussion directs the fidelity kernel toward data whose embedded states are "not too distant" in Hilbert space - the present study is, to our knowledge, the first end-to-end, referee-scored measurement of how much that guidance is worth: about two register sizes of practical headroom over the per-entry heuristic on this family, with tail statistics as the best-ranking feasibility instrument (Section 7). Rigorous shot-complexity analyses [8, 9] and shot-estimation heuristics [6] address the per-entry precision question; the end-to-end referee outcome under clustered structure is the gap this study fills.

6. The classical wall

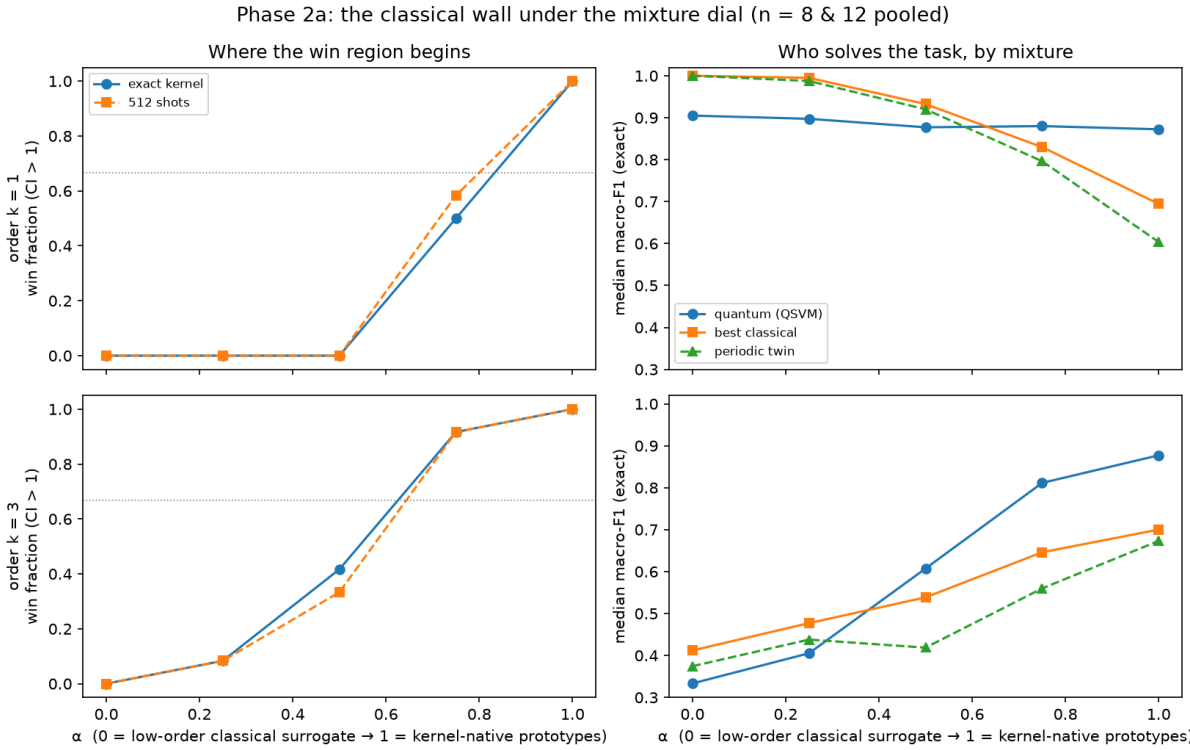

Figure 3 - The mixture dial. Left: win fraction against alpha (0 = pure sparse low-order labels, 1 = pure quantum-native labels) for surrogate orders k = 1 and k = 3, exact and 512-shot kernels. Right: median macro-F1 of the QSVM, the best classical variant and the periodic twin. At alpha = 0 quantum wins nothing; wins begin at alpha = 0.75. On k = 1 labels the QSVM stays near F1 0.9 and loses only to a perfect classical bar; on k = 3 labels it cannot represent the structure at all.

At alpha = 0 the win fraction is 0.000 at both estimation levels - the classical wall exists, and the referee holds. Win fractions rise monotonically in alpha, crossing the 2/3 threshold at alpha = 0.75 in three of four (order, size) combinations and at 1.0 for order-1 labels at 12 qubits. The alpha = 1 cells reproduce the Phase 1 results they duplicate (win fraction 1.0 both phases; median QAF 1.235 against 1.249) - the two families stitch into one consistent map.

Two findings inside this phase deserve their own statement:

Two ways to lose. On order-1 labels the QSVM's macro-F1 is essentially flat (0.91 to 0.87) across the entire dial while the best classical variant decays from a perfect 1.00 - at low alpha the quantum model learns the task and loses only to an unbeatable opponent. On order-3 labels the QSVM cannot represent the sparse-triple structure (F1 0.33 at alpha = 0). A parity ceiling and a wrong geometry produce the same verdict and deserve different consequences; the report vocabulary should keep them apart.

Representability is not findability. The registered expectation that order-1 labels would be easier for classical than order-3 labels was **refuted**, and the reason is instructive: on identical order-3 datasets the twin's greedy frequency search finds the three sparse triples on some train splits (F1 0.87) and misses them on others (F1 0.44), while the quantum model never learns them at all. Only the order-1 row is solved-classical territory in the strict sense (best classical >= 0.85 in every such cell); order-3 at alpha = 0 is partly no-man's land where nobody wins and classical merely stays ahead. The gap between what a model class can represent and what its search reliably finds cuts both ways - the same distinction that Section 7 documents for our own screening instruments.

7. Calibrating the instruments

The study's 124 datasets, each with a known referee outcome, form a labeled benchmark for the QML RQP's own pre-run instruments - a calibration the instruments had never had.

Verdicts failed. The fast assessment's quantum-territory checklist returned "classical territory indicated" on 83 of 84 quantum-native datasets - including all 79 island members - and returned the same answer at every mixture level, including on all eight pure quantum-native winners. Every suitability-rating class had median win fraction 1.0; as a discriminator the rating carries no signal on this family. The one reliable categorical signal is one-sided: the joint-spectrum probe's rare "dense joint spectrum" detection fired on six datasets, all island members (precision 6/6, recall near zero). The structural cause mirrors Section 6: the probes are built from order-<= 3 trigonometric designs, and a nearest-prototype boundary under an entangling map is outside their representable class - absence of probe signal is not evidence of classical territory.

Measurements succeeded. Scored as continuous rankers against island membership at 4,096 shots, the configured kernel's own Gram statistics lead the table: effective rank AUC 0.87 (bootstrap 95% interval 0.77-0.96), off-diagonal tail 0.86 (0.76-0.95) - rising to 0.97 (0.93-0.99) and 0.96 (0.92-0.99) against 512-shot membership, where feasibility is decided. The deep assessment's existing diagnostics rank well too, with wider intervals (concentration spread 0.83 (0.70-0.95), proxy advantage factor 0.77 (0.53-0.96), probe delta 0.75, kernel-target alignment 0.72 (0.57-0.85)) even though every categorical verdict layered on them miscalibrates: the concentration verdict declared "severe concentration" on every dataset from 10 qubits upward while trained kernels won everything through 12, and the proxy advantage factor never once said quantum-ahead across 79 winning datasets. One registered axis was dropped entirely: the entanglement-lift probe scored at or below chance (AUC 0.37-0.47). The classical-headroom axis - untestable on the quantum-native family, where classical never solves the task - was rescued by the mixture family (AUC 0.72 in-family) and serves as the gate against classical-solved data, with the no-man's-land cells as the standing reminder that high headroom alone can mean "hard for everyone", never "quantum opportunity".

The calibration is referee-robust: expanding the referee (Section 3.2) changed no arm's verdict, so island membership - and with it every AUC in this section - is unchanged.

The consequence is shipped. Since August 2026 the deep assessment computes the concentration of the configured kernel on a capped sample of the run's own data - bulk and tail, with a shots-to-resolve estimate keyed to the tail per Section 5 - and places the run on this study's benchmark map with the calibrated win-fraction bands printed next to it, upper-bound scope on the panel itself. Categorical verdicts on these quantities were not recalibrated; they were replaced by the measurements. The panel is kernel-methods-only, and the calibration is valid for this benchmark family: bands for other feature-map families and real-data anchors are re-derived, not transferred.

8. Hypothesis scorecard

Pre-registered outcomes, in the order registered. Refuted entries are results, not failures; three of them became the study's most useful findings.

Hypothesis	Outcome
H1 - classical wall at 1-2 prototypes (quantum-native family)	**Refuted.** Win fraction 0.86-0.95 even at one prototype; the twin never catches the family (Section 4).
H2 - shot wall co-located with the 3/sqrt(S) crossing	**Half-confirmed.** The wall exists and is budget-ordered; the per-entry criterion misplaces it by about two register sizes (Section 5).
H2' - exact-kernel erosion from finite training data	**Supported at 16 qubits.** Exact arms fall to 0.67 win fraction; shots steepen rather than replace the erosion.
H3 - non-empty island at 4,096 shots, moderate sizes	**Confirmed, understated.** 32 of 32 qualifying cells (Section 4).
H4 - screening gates conservative (high precision, low recall)	**Confirmed in the extreme.** Checklist recall about zero on winners; dense-joint detection precision 6/6 (Section 7).
P2-H1 - no wins at alpha = 0	**Confirmed.** Win fraction 0.000 (Section 6).
P2-H2 - monotone in alpha; order-1 crosses earlier than order-3	**Monotonicity confirmed; ordering refuted** - the twin's findability gap reverses it (Section 6).
P2-H3 - classical headroom discriminates once classical territory exists	**Partially confirmed.** AUC 0.72 in-family, 0.686 pooled (Section 7).
P2-H4 - gates correct on their home turf (alpha = 0)	**Refuted for the checklist** (constant across alpha); the suitability rating showed real one-sided signal (Section 7).
R1 - the island survives a referee expanded by tuned RBF-SVM and MLP	**Confirmed.** Zero verdict flips in 756 arms; win maps unchanged to three decimals (Section 4).
R3 - tail ablation: keep-top preserves F1, destroy-top collapses it	**Refuted in the strong form.** On exact kernels either channel alone retains most performance; the tail's advantage is specific to the shot regime at the largest sizes (Section 5).

9. Limitations

- **Upper bound by construction.** All labels are synthetic and quantum-native or mixtures thereof; no claim about real-world data follows from the winning region. Our own real-data benchmarks to date land in classical territory - which the instruments of Section 7 are built to detect before training.
- **One feature-map family.** zz-like embedding, two repeats, no input scaling, 1,000 rows, 60/20/20 split. Bandwidth scaling, other maps and row budgets are registered extensions, not results. Calibration bands are family-specific.
- **Measurement statistics, not hardware.** The shot model is per-entry binomial sampling with PSD projection; device noise (decoherence, readout error, drift) is a separate ladder - measured for QAOA in the companion noise-ladder study, not here.
- **Six replicates per cell.** Win fractions are counts over 2 generator x 3 split seeds; seed spread is reported wherever a headline number is quoted. The membership labels used for instrument AUCs are imbalanced on the quantum-native family (79 of 84); the 512-shot membership (69 of 84) and the pooled set (94 of 124) show the same ordering.
- **Kernel methods only.** A pilot on the variational classifier exists (Section 11) but supports no map-level claims; its study follows separately.

10. Reproducibility

The complete study - locked design with registered amendments, generator and referee engine, runner scripts, raw per-arm results (JSONL), analysis scripts, generated tables and figures - is committed under **studies/advantage-island-2026-08/** in the qubit-lab repository. Every number in this document traces to **tables*.generated.json**, regenerated from the raw results by the committed scripts. The pipeline is deterministic end to end: the production re-run of the demo benchmark reproduces the study values to four decimals, and the benchmark artifact shipped with the QML RQP (**island_reference_v1.json**) records its source commit. Referee, training path and feature maps are the shipping product's own code, not a re-implementation.

11. Outlook

A pilot arm on the variational quantum classifier initially suggested that the island belongs to the kernel method alone; our pre-specified verification protocol - a positive control plus systematic probes of the product's own training levers, with decision rules stated in advance - overturned that conclusion within a day, reaching statistical parity with the QSVM at 12 qubits under the product's published sample configuration and identifying training budget, not physics, as the binding constraint. That lesson - a fair run is part of the method - became the methodology of the companion study, The Advantage Island of Variational Classifiers (August 2026), whose measured answer completes the pair: a monotone, size-gated crossover between the two methods' territories, with each method failing for different reasons and screened by different instruments. A second follow-up places real financial datasets on the measured region: the search problem that Section 1 promises decision-makers, run in earnest.

Appendix A - The pre-specified design, amendments and deviations

The full locked design (DESIGN.md, 2026-08-27, with registered amendments of 2026-08-27/28) is committed alongside the raw results; this appendix records its skeleton and every deviation.

On the mechanism's scope and limits: we deliberately say pre-specified, not pre-registered. The design was locked in writing before any grid run, but its first public push postdates the first grid, so registration-grade chronology cannot be proven for this study; the companion study closes that gap with a publicly pushed, tagged design that provably precedes its runs, and future studies will add an independently timestamped registration before execution. Two calibration cells ran before the lock to size the budget and are re-run by the grid.

A.1 Registered phases

- **Phase 1 (locked 27 August, run same day):** the 756-arm grid of Section 3.3; hypotheses H1-H5; analysis plan fixed with the lock (win maps, concentration overlays, logistic main effects, seeds as spread, negative results printed).
- **Phase 1b/1c (amendment registered 27 August after Phase 1 analysis, before any calibration run):** fast- and deep-assessment stacks on all study datasets; candidate instrument axes and their orientations fixed a priori; promotion rule AUC >= 0.70.
- **Phase 2a (amendment registered 27 August, before any Phase 2 run):** the mixture family of Section 3.1, its grid, hypotheses P2-H1..P2-H4, and the generator validity gate (best referee variant >= 0.85 macro-F1 at alpha = 0 on a smoke cell, else revise and record before the grid).
- **Phase 1e (amendment registered 28 August after external technical review, before any reinforcement run):** referee expanded by tuned RBF-SVM and MLP with both grids re-run in full (original results retained separately); selection-aware panel-max bootstrap per arm; the tail/bulk ablation with its prediction stated in advance; PSD-projection sensitivity; cluster-bootstrap uncertainty for the logistic summary; and the wording upgrades of Sections 2-5.

A.2 Deviations, disclosed

1. **Calibration smokes before the lock.** Two timing cells (8 and 12 qubits) ran before the Phase 1 lock to size the budget and motivated the shots axis; one engine sanity cell ran pre-lock. All were re-run by the grid; none are reused.
2. **Validity gate fired.** The first low-order surrogate (a dense 2n-term random trigonometric polynomial) produced noise territory - best referee variant F1 0.55, quantum 0.41 - and was revised to the sparse three-term form; gate re-passed at twin F1 1.00 (order 1) and 0.87 (order 3). Recorded in the design before the Phase 2a grid ran.
3. **Membership substitution in the pooled instrument analysis.** Phase 2a dropped the 4,096-shot arms (Phase 1 had shown them statistically indistinguishable from exact through 14 qubits); the pooled AUC therefore uses exact-kernel membership for mixture datasets and 4,096-shot membership for quantum-native datasets, as registered in the amendment.
4. **Phase 1's fast-assessment stack ran on all 84 datasets** rather than the registered fallback subset (probe cost stayed below the threshold that would have triggered the subset).
5. **Instrument-panel workaround.** The offline assessment harness sets a display-only training-iterations field that the recommendation formatter requires; the value affects no measurement. The underlying defect was filed and fixed in the product separately.
6. **The shot-regime ablation condition was added after the exact-kernel ablation results were seen.** The pre-stated ablation covered exact kernels; its refutation (both channels redundant) motivated the additional condition - ablate the true kernel, then shot-sample - which was run once, unmodified, and is reported as stated. Readers should weight it as a follow-up probe, not a pre-specified test.
7. **The "advantage island" was measured against an eight-variant referee first and a ten-variant referee second** (Phase 1e); since no verdict changed, the document reports the ten-variant runs as primary and the original results remain committed for comparison.

A.3 Verification protocol (pilot arm)

The variational-classifier pilot and its verification (positive control reproducing a certified production run to four decimals; lever probes with pre-stated decision rules) are registered in the same design file as Phases 2b/2b-V. Because their outcome re-scoped the study to kernel methods, their results appear here only as Section 11's outlook; the raw records are committed with the rest.

Revision 1.1 (August 2026): presentation-level revision - question-format management summary, bootstrap confidence intervals added to the instrument calibration, pre-specification note moved to this appendix, companion-study cross-reference. All measurements and verdicts are unchanged from the first release.

References

1. S. Thanasilp, S. Wang, M. Cerezo, Z. Holmes, Exponential concentration in quantum kernel methods, Nature Communications 15 (2024); arXiv:2208.11060.
2. H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, J. R. McClean, Power of data in quantum machine learning, Nature Communications 12, 2631 (2021); arXiv:2011.01938.
3. J. M. Kübler, S. Buchholz, B. Schölkopf, The inductive bias of quantum kernels, Advances in Neural Information Processing Systems 34 (2021); arXiv:2106.03747.
4. R. Shaydulin, S. M. Wild, Importance of kernel bandwidth in quantum machine learning, Physical Review A 106, 042407 (2022); arXiv:2111.05451.
5. A. Canatar, E. Peters, C. Pehlevan, S. M. Wild, R. Shaydulin, Bandwidth enables generalization in quantum kernel models, Transactions on Machine Learning Research (2023); arXiv:2206.06686.
6. A. Miroszewski, M. Fellous-Asiani, J. Mielczarek, B. Le Saux, J. Nalepa, In search of quantum advantage: estimating the number of shots in quantum kernel methods, arXiv:2407.15776.

7. J. Mancilla, T. Tagliani, The Fourier Wall, arXiv:2607.15815 - the periodic-probe methodology of the referee's twin follows this work (independent implementation inspired by the paper; no affiliation).
8. Y. Liu, S. Arunachalam, K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nature Physics 17 (2021).
9. G. Gentinetta, A. Thomsen, D. Sutter, S. Woerner, The complexity of quantum support vector machines, Quantum 8, 1225 (2024); arXiv:2203.00031.
10. qubit-lab.ch, The Advantage Island of Variational Classifiers (August 2026) - the companion method study of the variational side; referee and mixture generator are shared. Published on qubit-lab.ch/evidence.

Disclaimer

This document reports a method study on synthetic benchmark data with a rapid quantum prototyping tool by qubit-lab.ch. Its figures are unedited outputs of the described runs, intended for technical review, education and structured discussion. The reported winning regions are a best-case benchmark envelope measured on labels constructed to favor the quantum method; they are not financial advice, not a performance claim for real-world datasets, and not a hardware result. qubit-lab.ch is not affiliated with the authors of the cited references.