AN ESTIMAND-SEPARATED AUDIT OF SIMULATED VARIATIONAL QUANTUM FEATURE MAPS ON TABULAR DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Variational quantum circuits are widely studied as feature maps for classical data, yet benchmark conclusions disagree, in part because they depend on the comparator, the objective, and the resources charged, and because evaluations pool three distinct questions: whether circuit features improve a task model, whether a classical student can reproduce a trained map, and how circuits and classical maps compare when trained directly on the task. We audit the three separately for one angle-encoded circuit family, primarily at $q{=}6$ under analytic-statevector simulation, on tabular benchmarks, binding each result to a stated cohort, comparator, and decision rule, and find that for this family and regime the observed gains trace to the classical front end and its nonlinear expansion, with no isolated gate-angle contribution detected. As augmentation, none of 28 circuit-versus-classical transform contrasts clears the benchmark-derived rule, and the one margin over raw features that does is not confirmed on an untouched cohort. Under distillation, four of 28 classical students reproduce a cached trained map, but none remains cheaper once generating and using the teacher's outputs are both charged, so reproduction is not replacement. In head-to-head training behind the same front end, screening rejects seven of 56 contrasts, none placing a circuit above a width-matched classical map. A paired optimizer control isolates the training gain at $+1.1$ pp, below the 2.0 pp criterion, and training the classical compression alone recovers most of it. The record supports no circuit-free replacement, hardware or end-to-end saving, or practical advantage.

1 INTRODUCTION

Variational quantum circuits are widely studied as feature maps for classical data (Havlíček et al., 2019; Schuld & Killoran, 2019; Cerezo et al., 2021), yet benchmark conclusions disagree (Bowles et al., 2024; Schnabel & Roth, 2025; Kakavand et al., 2026), in part because they depend on comparator, objective, and resource choices (Bowles et al., 2024; Schnabel & Roth, 2025; Schuld & Killoran, 2022; Gundlach et al., 2026). Few empirical evaluations separate three distinct quantities for a fixed variational-circuit readout family under protocol-specific thresholds: the utility of circuit features as augmentation, the reproducibility of an already-trained map by a classical student, and teacher-free task performance under a controlled, training-target-informed front end. Here and below, "teacher-free" means circuit-output-teacher-free, not target-blind. We provide that case-study audit — not a new circuit or general dequantization algorithm — and its ledger records target quantity, cohort status, comparator, rule, lifecycle, and permitted inference. The separation is an instrument exercised on this one family, not a contribution offered for reuse (Section 7). Three primary estimands cover augmentation, cached-map reproduction, and circuit-output-teacher-free task performance; Table 1's confirmation, optimizer, and width rows are supporting designs. The record's answer is one sentence: for this family and regime, the observed gains trace to the classical front end and its nonlinear expansion, and no isolated gate-angle contribution is detected. The separation matters: the circuit's mean margin moves by -49.17 pp when the comparator pool is widened, and its Nyström contrast falls from $+48.40$ to $+0.72$ pp between the 17- and 13-dataset cohorts (Section 5).

Two data protocols support three evaluation lanes: the augmentation protocol fits a linear estimator on $[x, f(x)]$ with a fixed-penalty primary analysis and tuned-alpha sensitivities, while the controlled-

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

front-end protocol supports teacher-dependent map distillation, in which classical students learn the circuit's width-23 outputs alone, and circuit-output-teacher-free task screening, in which circuits and classical comparators train directly on the task behind the same front end. The principal regression cohort contains 17 datasets with seeds averaged within dataset; circuit width is $q{=}6$, with a separate cost sweep to $q{\leq}12$, under analytic statevector simulation. The benchmark datasets were reused; a separate 12-dataset cohort, mechanically selected and scored once, provides the untouched confirmation test. The arm-level cached-map clauses and augmentation rule were recorded before their runs; the operational width-scaling aggregation was not. The 2.0 pp augmentation criterion is a benchmark-derived joint rule on mean, median, and win–loss record, not a universal threshold (Section 3).

Terms. A *lane* is one evaluation question with its own protocol, an *arm* one method configuration within it, a *cohort* the dataset roster a rule is judged on, and a *clause* one condition of a decision rule. In the task screen the *bar* is the best width-matched classical arm chosen on inner training folds, and the two *point hurdles* are beating C0 and the bar. T0 and T1 are fixed-penalty heads on unstandardized and fold-standardized designs; T2 and T3 tune the penalty under the same two treatments. Results read mean/median margin in pp with a win–loss–tie record, e.g. $+5.05/+2.54$ pp, 11–1–1.

Concretely:

- **Augmentation** (Section 5). The fixed-head rule is applied to seven transform arms under each of two readouts and the tuned-head rule to 14 Lasso contrasts; on the decision cohort none of the 28 clears the benchmark-derived joint rule. Separately, the fixed Lasso-T1 circuit-versus-raw-Original contrast clears it ($+5.05/+2.54$ pp, 11–1–1), as do the tuned-head sensitivities (T2, T3; Table 1); neither is untouched confirmation, and the one-shot untouched-cohort test does not confirm the rule (one catastrophic dataset defeats the mean clause; median and record pass, $+4.67$ pp, 10–1–1). A same-run paired control is positive on its observed roster; under a non-step-matched component ablation, compression-only training recovers most of the observed bundle effect and no isolated gate-angle gain is detected.

- **Teacher-dependent map reproducibility** (Section 4). The four firing arms — C0 with a degree-3 ridge or fitted random Fourier features (RFF), and C-edge and C-GraphRZZ with fitted RFF — reproduce their map-only readouts with held-out R^2 of 0.956–0.982 and retain 99.6–100.1% of downstream scores. These results concern distillation of the tested trained readouts, not circuit-free replacement; the separate q-scaling stop does not fire.

- **Circuit-output-teacher-free task screening** (Section 6). The preregistered five-seed screen uses 13 datasets per task and an inner-fold-selected bar; seven of 56 corrected contrasts reject, none placing a circuit above a width-matched classical arm, and two reuploading arms clear both point hurdles in the regression practical lane on intervals the reduction calls inconclusive. The raw-XGBoost and natural-width cubic ceilings remain outside the width-matched decision family, and familywise equivalence does not follow.

Table 1 consolidates the lane status, cohort, comparator, rule, result, and permitted inference; the appendices preserve the full record.

Supporting diagnostics qualify rather than extend these three results. Three data-reuploading measurements are correlated readings of one trained readout, and an exploratory gate-budget ablation cannot resolve its margin against its threshold (Section 6; Appendices Q, U, and O). Across four defensible absorption conventions the quadratic span explains 94.97–97.28% of held-out readout variance, and the registered C-edge versus learned-frequency-RFF contrast is equivalent under the ± 1 pp margin, with equality rejected (Section 6). A one-seed synthetic screen nominates high-order sign interactions, but no dataset-level contrast is significant and the nominated direction fails its 17-target transfer test. The record retains decision rules, named misses, reused-data status, and cohort limits (Sections 3 and 7).

2 RELATED WORK

Quantum feature-space and fixed-parameter/kernel methods developed in parallel (Wilson et al., 2018; Havlíček et al., 2019; Schuld & Killoran, 2019; Huang et al., 2021), but classical-data conclusions remain comparator- and protocol-sensitive. Schnabel & Roth (2025) cover five families and 64 mostly controlled or generated instances. In Kakavand et al. (2026), 29 initial tests each use five paired observations; the two-sided exact Wilcoxon signed-rank floor is 0.0625, so non-significance is not parity. Comparator choice, objectives, and resource barriers likewise matter (Bowles et al., 2024; Schuld & Killoran, 2022; Gundlach et al., 2026), while concentration theory motivates explicit width checks (Thanasilp et al., 2024).

Task settings and encoder roles separate adjacent empirical precedents: regression spans synthetic approximation (Panadero et al., 2024), controlled kernel studies, and task-direct tabular variational quantum circuits (VQC). Belis et al. (2024) use task-guided trainable compression; Su et al. (2026) use a modular VQC without a trainable classical encoder; QuFeX is evaluated at a U-Net bottleneck for image segmentation (Jain & Kalev, 2026). Boosted trees, established tabular learners, and classical feature generation are practical baselines (Grinsztajn et al., 2022; Zhang et al., 2023); our width-matched MLP and smooth bases are mechanistic controls, and same-protocol practical coverage remains incomplete. This distinction motivates circuit-output-teacher-free screening.

Spectral structure, data re-uploading, and trainability analyses provide adjacent circuit-design precedents (McClean et al., 2018; Pesah et al., 2021; Schuld et al., 2021; Pérez-Salinas et al., 2020). Landman et al. (2023) directly approximate VQCs with architecture-informed RFF, and Sweke et al. (2025) give necessary and sufficient conditions for efficient RFF dequantization in regression. Our fitted RFF instead post-hoc fits cached trained-readout targets; it is not a new RFF mechanism. Sahebi et al. (2025) bound RFF–quantum generalization gaps for regression and classification with quantum-neural-network and kernel architectures, whereas our audit supplies no dequantization bound and fits cached outputs from particular trained readouts. Schreiber et al. (2023) give guarantees for direct surrogate optimization. Nair & Ferrie (2025) chain local quantum and classical surrogates to cut inference cost after quantum training; our builder-fit audit instead conditions on cached teacher outputs and claims no end-to-end or amortized saving. Cerezo et al. (2025) simulate provably plateau-free families, sometimes after quantum data acquisition; and Gil-Fuster et al. (2025) establish that trainability and non-dequantizability coexist, which is not practical utility.

For public tabular data, Mancilla & Tagliani (2026) propose a simulator-free spectral screen with matched quantum/classical tests on nominated tasks. Ours is narrower — it certifies no target spectrum, uses no untouched tasks, and yields no practitioner rule — and that estimand difference prevents either study from validating the other.

These benchmark studies do not jointly report incremental augmentation, trained-map reproduction, and circuit-output-teacher-free task performance with explicit question-specific arm and cost protocols (Bowles et al., 2024; Schnabel & Roth, 2025; Kakavand et al., 2026). We evaluate them in separate lanes under two protocols, with lane-specific cohorts, accounting, and unequal power. This case-study organization neither validates a general method nor implies a joint recurrence.

3 SETUP, THE TWO PROTOCOLS, AND THE DECISION RULES

Pipeline and circuit family. Every supervised or preprocessing component is fitted on the training partition of each split and seed before it is applied unchanged to held-out rows (Figure 1). The regression augmentation transform fits three complete pipelines independently, with child seeds s, $s{+}1000$, and $s{+}2000$: mutual information (MI) preselection and selected-column quantile transformation are fitted first, after which supervised compression and circuit parameters are optimized jointly. Each circuit emits $4q{-}1$ features and the three members $3(4q{-}1)$: 23 and 69 at the primary $q{=}6$. The controlled comparisons instead evaluate one width-$(4q{-}1)$ map per arm. Every arm reconstructs the same frozen, training-only MI–quantile–principal-component-analysis (PCA)–affine front end, with matching split, compression, and coordinate hashes; MI reads y_{train}, so this front end is target-informed, fitted before and not co-trained with the circuit. Held-out coordinates are not clipped, no controlled arm uses the three-member concatenation, and Appendix B specifies the augmentation initializations, subsampling, and both front ends. The audited family is the authors'

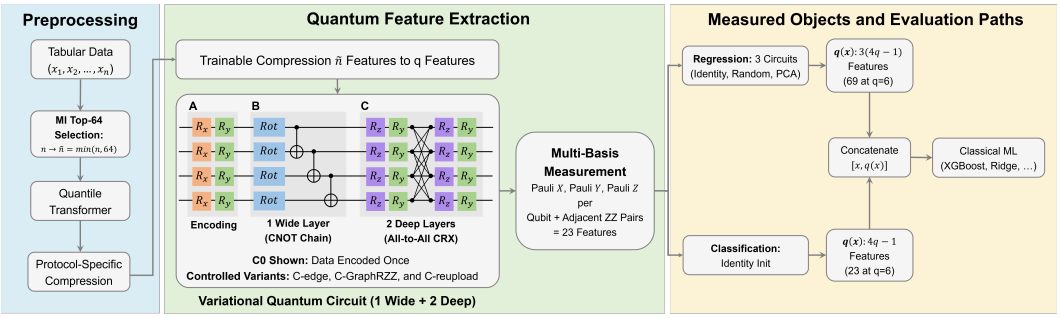

Figure 1: Architecture overview. Mutual-information top-64 selection precedes the quantile transform, protocol-specific compression, the circuit, and multi-basis measurement, which emits $4q-1$ features per circuit, 23 at the primary q=6. The center circuit is C0, the 1W+2D ansatz with data encoded once. The register is drawn at q=4, but the unlabeled 23 in the measurement box refers to q=6. The three-circuit regression stack, the original-feature concatenation, and the "Trainable Compression" box are augmentation-specific; controlled arms instead enter one map from the frozen PCA–affine front end. Panels A–C: Appendix C.

own prior design, audited here rather than extended; Appendix J keeps the earlier record under its original labels.

The controlled family contains four compound variants: C0, the 1W+2D ansatz; C-edge, with edge-specific deep angles; C-GraphRZZ, which selects and freezes a training-only sparse graph before applying graph-ZZ gates; and C-reupload, with block-specific encodings but tied variational tensors. All emit $4q-1$ features, but they are neither parameter- nor compute-matched and therefore do not isolate single causal mechanisms. The augmentation +Reupload construction is different because it reuses the same encoding parameters at each insertion. Primary cells use q=6; sensitivity and cost experiments vary q. Appendix C gives the gates, topology, parameterization, and parameter counts.

Controlled circuits maximize centered kernel alignment (CKA), not the augmentation loss. Regression uses a standardized scalar label factor and classification a class-balanced one-hot factor; a disjoint inner holdout selects the checkpoint. Appendix B gives the objective, optimizer, row caps, and checkpoint schedule.

Data, scoring, precision. The augmentation benchmark, the 17-dataset controlled regression lanes, and the governing task screen use seeds $\{42, 123, 456, 789, 2024\}$. The screen covers 13 regression and 13 classification datasets, with standardized root-mean-square error and balanced accuracy as its respective primary metrics. The augmentation grid uses benchmark-wide fixed defaults, except that Tuned MLP performs a 20-trial inner search per dataset. Its "overall" column maximizes the outer-test score over downstream estimators and is descriptive; a fixed-readout view takes priority on disagreement. The grid caps each dataset at 10,000 rows before a 70/30 split, separately from the at-most-5,000-row circuit-fitting subsample. Controlled lanes retain their own sampling. Eight of 12 design settings were selected using the benchmark, and nominally repeated fits differ by up to 1.32 pp; none of the 37 datasets is untouched confirmation material. A confirmation cohort of 12 OpenML-CTR23 regression datasets was selected mechanically and version-pinned before any score existed, and scored once (Section 5). Full settings and provenance appear in Appendices K and L. Section 5 regenerates and seals the benchmark splits before scoring. Every scored cell uses PennyLane 0.38.0's `default.qubit` analytic-statevector backend (Bergholm et al., 2018); a finite-shot study in Appendix F is the stated exception.

The two protocols. The augmentation protocol fits a supervised compression per arm, trains the circuit with a hybrid loss, and scores every transform as $[x, f(x)]$ at its native output scale. The controlled protocol freezes a deterministically reconstructed train-fitted MI–quantile–PCA–affine front end before fitting any circuit, requires its split, compression, and coordinate hashes to match across arms, aligns models by CKA, widens the comparator pool, and fits downstream heads to each map without original-feature concatenation. Thus the augmentation protocol asks what the

map adds with the original features retained, whereas the controlled protocol asks whether the map output alone stands where a classical transform would. The decisive augmentation settings are q=6, at most 100 training steps, three independently fitted regression members, concatenated $[x, f(x)]$ output, and fixed-head readout priority when it disagrees with the descriptive outer-test maximum; Table 7 gives the full specification.

Claim and protocol registry. The three primary evidence lanes have different outputs and estimands; no joint effect is formed. Of Table 1's six evidence rows, confirmation and optimizer attribution support augmentation, while finite width supports cached-map reproduction. The augmentation benchmark is adaptive and uses descriptive best-of-estimator and primary fixed-readout views. The complete 17-dataset record is reported before the outcome-informed 15- and 13-dataset sensitivities. The tuned-alpha treatments are head-only sensitivities conditional on representations already fitted on the full outer-training split; the standardized treatment additionally standardizes within head-selection folds. They are not fully nested representation tuning and remain outcome-informed sensitivities rather than untouched confirmation; the registered transform family excludes Original, whose tuned-head contrast is reported as a separate sensitivity (Section 5). The span lane's reporting convention was fixed after its sweep. Cached-map distillation separately measures readout R^2, fixed-readout task margin, and lifecycle-bounded cost without established per-run prospective enforcement. The task screen (Section 6) is teacher-free only with respect to circuit-output targets because MI still reads y_{train}; its governing five-seed reduction can be rerun, but not its producer.

Decision rules and analysis conventions. Claims are tracked by an orthogonal checklist: A0, a circuit was used; A1, a held-out gain over a named comparator set under a named preprocessing state; A2, sample efficiency; A3, practical advantage; A4, computational advantage. A1 appears only with comparator set and standardization state attached; A2–A3 are not measured here; A4 is governed by the stop condition of Section 4. For the augmentation decision only, the 2.0 pp threshold is a repeatability-based analysis threshold, rounded above the benchmark grid's median across-seed deviation of 1.35 pp per dataset but below its 2.77 pp mean; it is not an elicited deployment-utility threshold. An augmentation contrast clears it exactly when its mean exceeds 2.0 pp, its median exceeds 2.0 pp, and wins outnumber losses. The rule has no calibrated familywise operating characteristic, and failure to clear it is not evidence that all meaningful effects are absent. Equivalence requires the complete dataset-clustered 95% interval to lie inside ± 2.0 pp. An interval that excludes zero supports superiority or inferiority; one that covers zero and escapes the margin is inconclusive, not parity. The finite-width rule was amended after feasibility failures and is treated as operationally non-confirmatory; it is excluded from any claim that decision rules preceded their corresponding runs. The unit is the dataset: seeds fold first, and interval estimates are dataset-clustered and labeled with their statistic. The task screen and augmentation lane report lane-specific power or minimum-detectable-effect quantities; cached-map distillation follows its own three-clause stop. Regression and classification are not pooled for current inferential verdicts; any 37-dataset cross-task mean is descriptive and is not a common-scale effect estimate. Dataset-bootstrap intervals quantify sensitivity to reweighting the observed dataset roster; they do not target a task superpopulation. The protocol record includes amendments and exploratory or broken records, so its size does not establish prospective enforcement; Appendix G records the available checks and the task-screen and distillation gaps in pre-run validation and sources.

4 CACHED-MAP DISTILLATION AND LIFECYCLE-BOUNDED COST

Scope and accounting. In this conditional distillation audit, students fit cached train-row outputs of an already-trained circuit and are evaluated on the sealed test split; this is not a teacher-free replacement test (Section 6). Ledger-declared *builder-fit* accounting charges circuit fitting plus the shared downstream fit/test chain to the circuit, and student fitting plus that chain to the student; it assumes teacher outputs exist and excludes their generation, so it is neither marginal serving nor end-to-end construction cost. Recorded-component sensitivities are in Appendix O.

Arm-level rule and result. Each of 28 correlated variant–student arms is judged by three ledger-declared clauses: lower builder-fit cost, mean retention of at least 95%, and a nominal dataset-clustered 95% confidence interval (CI) for the student-minus-circuit downstream margin with lower endpoint at least -1.0 pp. Retention averages per-dataset student/circuit ratios over 16 usable

Table 1: Evidence rows under three primary estimands. Confirmation and optimizer rows support augmentation; finite width supports cached-map reproduction. Entries read mean/median margin (pp) with a win–loss–tie record; T1 is the fixed head on fold-standardized designs, T2/T3 the tuned head.

Evidence row and status	Design	Result and permitted inference
Cached-map reproduction *primary; reused*	*Cohort:* 17 regression datasets (retention over 16 usable). *Comparator:* 7 students per variant. *Rule:* lower builder-fit cost; retention $\geq 95\%$; CI lower bound ≥ -1 pp.	*Result:* 4/28 pass; teacher generation alone leaves 4/4 cheaper, marginal use alone 1/4, both 0/4. *Inference:* cached-readout reproduction under builder-fit accounting.
Augmentation *primary; adaptive; reused*	*Cohort:* 17 recorded; outcome-informed 13 judged. *Comparator:* 7 transforms; raw Original separate. *Rule:* mean and median > 2 pp; W>L.	*Result:* transforms 0/14 fixed, 0/14 tuned (Lasso); raw Original clears fixed Lasso-T1 ($+5.05/+2.54$ pp, 11–1–1) and tuned T2/T3 ($+5.77/+2.77$; $+6.74/+3.99$ pp). *Inference:* raw margin holds under head tuning here; not confirmation.
Untouched confirmation *augmentation support; prespecified*	*Cohort:* 12 CTR23 datasets, selected and version-pinned before any score; scored once. *Rule:* unchanged joint rule vs. raw Original, fixed Lasso-T1.	*Result:* not confirmed — mean -21.95 pp (one catastrophic dataset); median $+4.67$ pp, 10–1–1 pass; paired control $-13.85/+1.24$, 10–1–1. *Inference:* confirmation failure; the mean clause's catastrophic-cell vulnerability is measured.
Task screening *primary; circuit-output-teacher-free*	*Cohort:* 5 seeds; 13 datasets per task. *Comparator:* inner-fold width-23 bar; XGBoost and cubic ceilings. *Rule:* two point hurdles; 4 Holm-14 families.	*Result:* 7/56 adjusted rejections, none placing a circuit above a width-matched classical arm; 2 reuploading arms clear both hurdles in one stratum on inconclusive intervals. *Inference:* corrected ordering in the classical direction only; no familywise equivalence; front end target-informed.
Paired optimizer control *augmentation support; reused*	*Cohort:* amended 13-dataset decision cohort. *Comparator:* zero-step and group-frozen same-circuit arms. *Rule:* descriptive 2 pp convention.	*Result:* bundle $+1.137/+0.639$ pp; gate-angle-only $+0.050/0.000$ pp (6–6–1); compression-only $+1.027/+0.776$ pp (10–2–1). *Inference:* compression-only recovers most of the bundle; no isolated gate-angle gain is detected, step-matched or not (Appendix E).
Finite-width reproduction *cached-map support; non-confirmatory*	*Cohort:* amended 6 datasets. *Comparator:* selected students. *Rule:* 3-clause stop.	*Result:* no variant stops. *Inference:* no scaling, timing, or practical claim.

datasets after the ledger-declared denominator floor; margin and cost use all 17 datasets, with cost comparing mean seconds rather than per-dataset ratios. Table 8 gives the clause accounting (Appendix O). Builder-fit cost passes for all 28 arms, so the nominal CI clause is the tightest.

Within this cached-output conditional-distillation lifecycle, four arms meet all three marginal clauses — C0/degree-3 ridge, C0/fitted RFF, C-edge/fitted RFF, and C-GraphRZZ/fitted RFF; none is C-reupload. A post-hoc one-sided exact sign-flip Holm-28 screening sensitivity retains exactly those four, but its sign-symmetry assumption is unverified and it is not registered familywise inference. Charging teacher-output generation to one-batch student construction leaves all four cheaper, by roughly 117, 59, 56, and 50 times; under marginal-use accounting, which treats circuit fitting as sunk, three of the four are no longer cheaper, and none is once both changes apply. Full reproduction and accounting results are in Appendix O.

Finite-width audit. Both cost clauses pass in all 24 dataset–variant cells under the amended six-dataset selected-policy cohort rule, but the reproduction clause prevents every variant-level stop, so the three-clause prediction fails for all four variants. Its timing is a post-hoc five-to-30-step extrapolation, neither measured end-to-end nor confirmatory; Appendix O carries the operational rule, judged grids, learner selection, timing definitions, and two unjudged partial curves.

5 WHAT FIXED LINEAR-READOUT MARGINS MEASURE

Protocol and reproduced arithmetic. The analysis uses regression, five seeds, a fixed Ridge head (α=1), and a fixed Lasso head (α=0.01). T0 fits each head to the arm's unstandardized full design matrix. T1 fits a fold-local standardizer on the complete per-arm design matrix $[X_{\text{raw}}, f(X)]$ on the outer training rows and applies it to the test matrix, changing raw as well as mapped columns; for Original, the matrix is the raw block itself. The T0 scores agree across the two preserved measurement sources (1,700 cells, maximum $|\Delta| = 0$), so the issue below is interpretation rather than score reproduction.

The ℓ_2 margin. On the full 17-dataset descriptive cohort, the fixed-Ridge circuit-minus-Nyström contrast is $+8.56/-0.19$ pp with a 7–10–0 record and dataset-clustered mean CI $[-0.97, +25.44]$. Its 18.87 pp approximate normal-power scale derives from the CI width, not the exact sign test or the joint rule, and the discordant summaries leave the contrast inconclusive. On SkillCraft the unsealed grid gives $+93.42$ pp against $+134.10$ pp sealed (Appendix P); removing SkillCraft changes the sealed mean from $+8.56$ to $+0.72$ pp.

The ℓ_1 comparator pools. After excluding OpenFE, whose transform is not row-local and exists on only 4 of 17 datasets, the evaluated initial pool contains five comparators (Original, PCA, Poly-2, RFF, Nyström); adding Poly-3, Trained RFF, Budget-matched MLP, and Wide MLP gives nine comparators and ten arms. On the full 17-dataset cohort the four added transforms score 54.7–61.4, three exceed the circuit, which ranks fourth of ten, and pool expansion moves the circuit margin by $-49.17/-3.61$ pp (mean/median, expanded minus initial); Table 9 reports the paired T0 contrasts.

Standardization and cohort sensitivity. The full 17-dataset descriptive cohort is the complete record; the outcome-informed 15-dataset sensitivity omits SkillCraft and California Housing, and the amended 13-dataset decision cohort additionally omits Kin8nm and Space GA and is not untouched confirmation. Score coincidences with Original at tolerance 10^{-9} fall under T1 to 1/17 for Nyström, RFF, and the circuit; these are coincidences, not identical-prediction claims (Appendix P). The Nyström T0 ladder is $+48.40/+1.16$/9–5–3 on the full 17-dataset descriptive cohort, $+1.65/0.00$/7–5–3 on the 15-dataset sensitivity, and $+0.72/0.00$/6–5–2 on the amended 13-dataset decision cohort, with the T1 rungs in Appendix P. The last two cohorts therefore show sensitivity to the failures that motivated the exclusions rather than independent confirmation.

On the outcome-informed amended 13-dataset decision cohort, the registered seven-arm non-PCA decision family is Poly-2, RFF, Nyström, Poly-3, Trained RFF, Budget-matched MLP, and Wide MLP. None clears the registered, benchmark-derived three-part rule under fixed Lasso-T1 or Ridge-T1; no displayed contrast equals 2.0 pp, so strict inequality changes no verdict. The largest residual is against RFF under both readouts: $+2.13/+1.02$/10–2–1 on Lasso-T1 and $+2.88/+1.39$/12–1–0 on Ridge-T1, the larger of the two, which passes the mean and record clauses and fails on the median. Original is the sole raw-feature baseline; its Lasso-T1 contrast, $+5.05/+2.54$/11–1–1, clears the three-part rule (the median clause fails under six of 13 single-dataset deletions; Appendix P) but establishes no practical advantage; tuned raw-feature boosting exceeds its Lasso-T1 score by $+1.7$–2.1 pp median on 14–15 of 17 datasets (Table 6). PCA's T1 estimate is numerically divergent; its mechanically favorable mean, $+8.64$ median, and 11–1–1 record yield no interpretable joint-rule decision. Table 9 contains T0 paired contrasts and arm-level T0/T1 scores; Table 10 records the complete seven-arm Lasso-T1 decision set separately from Original, PCA, and quantum kitchen sinks (QKS). Original is outside this transform decision family; its tuned-head contrast is reported below as a separate sensitivity, and failure to clear the joint rule does not establish that every meaningful effect is absent.

What the scale interventions establish. Figure 7 and T1 show that the fixed-penalty pipeline is scale-sensitive and its rankings are not invariant to the tested scale and standardization treatments; they do not isolate a unique mechanism, since uniform map scaling and full-matrix standardization also change effective regularization, anisotropy, covariance conditioning, and raw–mapped column interactions. The protocol defines T2 as inner-fold alpha selection without standardization and T3 as fold-local standardization plus the same selection; both are head-only sensitivities conditional on a representation fitted once on the outer-training split (Section 3). The locked sensitivity contains all 2,080 cells: 13 datasets, five seeds, eight arms, two heads, both treatments. None of the 14

Lasso circuit-versus-transform contrasts clears the joint rule. The largest T2 mean is $+1.05$ pp against RFF, with a $+0.74$ pp median; the largest T3 mean is $+1.94$ pp against RFF, with a $+0.86$ pp median. In a separate sensitivity outside this family, Original scored under the identical T2/T3 protocol clears the joint rule at $+5.77$/$+2.77$ pp (11–1–1, T2) and $+6.74$/$+3.99$ pp (12–1–0, T3), so neither the transform decisions nor the raw-baseline margin is confined to fixed α=0.01, though the T2–T3 shifts retain material preprocessing sensitivity. These outcome-informed runs are controlled sensitivities, not untouched confirmation; Appendix P gives all 14 contrasts, the Original sensitivity, and hashes.

Untouched-cohort confirmation. A one-shot test, prespecified in selection and run locks, scores the fixed Lasso-T1 circuit-versus-Original contrast on 12 OpenML-CTR23 regression datasets selected and version-pinned before any score existed. The rule is not confirmed: the mean clause fails at -21.95 pp on one catastrophic dataset where every mapped arm collapses under the unbounded-below R^2 metric (circuit -464.4; raw -28.3); the median and record clauses pass ($+4.67$ pp, 10–1–1). The paired control there gives -13.85/$+1.24$ pp, 10–1–1, not positive under its prespecified mean summary; the mean clause's catastrophic-cell vulnerability is thus measured, not hypothetical. Appendix P records the selection audit, hashes, the grid's full set of post-hoc fixed-head contrasts — one of which, Ridge-T1 against RFF, clears the rule — and the prespecified clipped-at-zero summary, which clears all three clauses ($+14.38$/$+4.67$ pp, 10–1–1).

The sealed same-arm block deletion is unstable across treatments rather than scale-proof; its sign-flipping mean and small medians establish no robust raw-block contribution (Appendix P).

Distinct untrained-pipeline and same-circuit controls. QKS is a post-hoc, two-seed, untrained random-circuit benchmark pipeline, not the circuit map with training switched off, and differs in circuit family, register sizing, circuit count, Z-only readout, output width, and front end. Its three contrasts favor the circuit on the amended decision cohort (Appendix P), but compare two pipelines and cannot isolate parameter training.

The same-run paired Lasso control sets the full joint optimizer against a zero-step control; on the amended 13-dataset decision cohort its gain, $+1.137$/$+0.639$ pp, is a full-bundle point estimate below 2.0 pp. In a five-arm ablation over the same 65 cells, gate-angle-only training is null ($+0.050$/0.000 pp, 6–6–1) and compression-only training recovers most of the bundle ($+1.027$/$+0.776$ pp, 10–2–1) under a non-step-matched protocol; Appendix E gives intervals, records, step counts, the step-matched rerun, and the Ridge control bundle. Every one of these effects is smaller than the $\approx$1.3 pp repeatability floor this paper measures for same-configuration reruns (Appendix L), which under that appendix's own policy would leave none of them attributable to a design setting. They are read here only because the contrast is same-run and paired: the arms enter the optimizer from initializations verified identical by hash, 260 equality checks across 185 distinct hashes in the control and 1,040 in the ablation, so the run-to-run variation the floor measures is differenced out rather than assumed away. That argument does not extend to any comparison in this paper that is not same-run and hash-matched.

6 Circuit-output-teacher-free task screening and secondary diagnostics

Task-trained models without circuit-output teachers. Classical task models train on the shared standardized compressed coordinates behind the same frozen training-target-informed front end. The governing record is a preregistered five-seed run over 13 datasets per task with fixed-head and eight-trial head-selection lanes and a width-23 bar selected on inner training folds (Section 3): 46 arms drop (16 in the width-unmatched class) and two survive. Seven of the 56 contrasts in the four Holm-14 families reject, each on a 12–1 record at p_{Holm}=0.0479: four place a circuit below the width-matched wide MLP (first-arm-minus-second -0.0246 and -0.0211 sd in regression, -1.30 and -1.78 pp in classification), one places C-GraphRZZ-residual below C0, and two place C0 above the trivial width-6 floor. None places a circuit above a width-matched classical arm. Two reuploading arms nevertheless clear both point hurdles (regression practical), beating C0 by $+0.0159$ and $+0.0144$ sd and the inner-selected bar by $+0.0032$ and $+0.0017$; both carry the mean while losing most datasets, both reach p_{Holm}=1.0 against the wide MLP on intervals the reduction calls inconclusive, and the registered falsification clause — both references beaten by more than the

registered smallest effect size of interest (SESOI) with an interval excluding zero — is not met. The eight-trial cap binds unequally on four heavy datasets, but only on the raw-feature XGBoost arm, so no circuit or contrast reference is capped. In the representation lane that ceiling sits 0.119 sd and 7.87 pp above the best width-matched circuit and stays outside the width-matched families; the natural-width cubic exceeds the best width-matched circuit on classification only. The screen establishes a corrected ordering in the classical direction but neither familywise equivalence, replacement, mechanism, nor a practitioner recommendation, and Appendix Q preserves the full record.

Readout-span diagnostics. Training-split ridge maps from quadratic bases to held-out circuit readouts give 94.965–97.280% span across four width and aggregation conventions; the largest value has a dataset-clustered interval $[96.08, 98.21]$ over 17 regression datasets, but its convention was selected after the sweep began, so the record is descriptive. In the separate controlled pipeline, C0's declared quadratic span is 91.58% $[88.36, 94.42]$, and all six variant–basis exact sign tests reject after Holm correction. For the registered C-edge learned-frequency focus, the difference, -0.401 pp $[-0.682, -0.050]$, lies within the registered ± 1 pp equivalence margin: equality is rejected while margin equivalence holds. These are teacher-dependent readout reconstructions, not task replacement, functional equality, or mechanism evidence; Appendix U preserves every convention, level, interval, record, and adjusted test.

Secondary diagnostics. The matched gate-budget, routed-count, hybrid, and correlation diagnostics have mixed or null outcomes: they neither isolate a mechanism, explain the small optimizer-bundle effect, nor change a primary verdict (Appendices Q, T, R, U, and E). Five reused synthetic geometries nominate a comparator-conditional property whose raw-dimension proxy sends both standardized heads opposite to prediction; no selection rule follows (Appendix S).

7 Limitations and scope

Inferential boundary. Primary task inference is limited to the analytic-statevector, q=6 angle-encoded family; wider q=4–12 reproduction and q=8 noise records are diagnostics. Hardware and practical advantage are unmeasured, the lanes share no effect, and the one-shot untouched mean fails on one catastrophic cell (fixed heads only). The finite-width audit reuses held-out data, extrapolates five-step timing, and stops no variant (Appendix O).

Design and attribution limits. Variants are compound and unmatched; the paired bundle point estimate is below 2.0 pp, and its gate-angle isolate is null under both non-step-matched and step-matched protocols; circuit necessity and component interactions remain unidentified. The audit design is evaluated only in this case study. Task screening, outcome-informed cohorts and head tuning, teacher-dependent spans, and cached-output accounting support no practitioner rule; the untuned and tuned boosting ceilings (Table 6) are descriptive, while deployment cost is unmeasured. The archive validates artifacts, not a clean reconstruction of every input and run; the manifest records regeneration boundaries without inferring unknown fields (Appendices E, Q, P, U, and O).

8 Conclusion

As augmentation, no fixed- or tuned-head circuit–transform contrast clears its rule; the raw-Original margin clears only in outcome-informed analyses; one-shot confirmation fails its mean clause. Under distillation, four of 28 cached-output arms meet marginal clauses, none cheaper once teacher generation and use are charged. In head-to-head training, screening rejects seven of 56 contrasts, none placing a circuit above a width-matched classical arm. The optimizer bundle is below 2.0 pp; compression-only training recovers most of it and no gate-angle isolate is detected, step-matched or not. For this family and regime, the gains trace to the classical front end and its nonlinear expansion; no broad causal, hardware, end-to-end, or practical claim follows.

Reproducibility statement

The anonymous `Supplementary.zip` contains the `qfx_v2` Python implementation and tests, protocol manifests, compact ledgers and reports, the ICLR experiment scripts and result summaries,

a machine-readable dataset provenance manifest, environment metadata, and a hashed inventory. It excludes large dataset containers, model arrays, pickle files, caches, lock files, and raw or failed logs. The dataset manifest distinguishes recovered, reconstructed, and missing status, records checksums where evidenced, and uses null rather than inferred values for unknown fields. An E8 fail-closed wrapper is tested in both pass and deliberate-mismatch modes. A retrospective E5 check finds all 2,720 scoped `e5_full` rows valid and successful, with no duplicates, missing cells, or unexpected family members; this establishes present completeness, not prospective enforcement of the protocol. The four firing arms of Section 4 — DQ-1 in the protocol record and the archive — are also evaluated by an explicitly post-hoc exact sign-flip Holm-28 screening sensitivity whose sign-symmetry assumption is unverified. The finite-width audit, DQ-2 in the same record, remains operationally non-confirmatory. Appendix O reports its post-hoc sequential single-worker timing sensitivity: 120 cells, 600 width points, and 600 five-step timing runs, with the five-step circuit-fit component scaled by six rather than a measured 30-step end-to-end run. Appendix G identifies exact noise and expressibility values that remain artifact-only because their compatible producing drivers are absent. For the task screen's governing five-seed record (run `e4v2_expand26_20260827`), the archive packages the preregistration, the producing driver, the ledger, and the reducer, but omits the required roster container: a clean extraction therefore reproduces the reduction exactly and cannot rerun the producer. The cached-map lane is likewise audit-only from the archive: trained circuit parameters and cached teacher outputs are excluded, so its 28 arms can be re-reduced from the retained ledgers but not regenerated. The superseded eight- and four-datasets-per-task predecessors are retained alongside it, and the original one-seed screen and the concatenation record remain artifact-only. The paired E7 control completed 65/65 same-run cells in the same recorded runtime environment; all 260 optimizer-entry equality comparisons pass across 185 distinct hashes, with 19,346 trained-arm steps versus zero control-arm steps. Its five-arm component ablation passes all 1,040 optimizer-entry equality comparisons and all frozen-group hash checks, and reproduces the control's trained and zero-step scores exactly. These are step counts, not timing evidence. Its analysis emitted a non-fatal SciPy binary-compatibility RuntimeWarning, so the environment is not represented as clean; the warning is retained in packaged analysis metadata.

ETHICS STATEMENT

This work uses public tabular datasets only; it involves no human subjects and no participant-facing procedures. Claims are graded on the explicit checklist of Section 3, under which each claim level requires its own evidence and the strongest level is withheld by the relevant protocol-specific stop conditions. The positive task margins are outcome-informed circuit-versus-raw-Original results on classical simulators with penalized linear regression readouts; they are not confirmation. None of the fixed-head circuit-versus-transform comparisons on the decision cohort clears the 2.0 pp joint rule under either readout, and none of the 14 tuned-head Lasso comparisons clears it, and neither summary of the same-run paired full-joint-optimizer-on versus zero-step contrast exceeds 2.0 pp; failure to clear this uncalibrated familywise convention does not establish the absence of all meaningful effects. The circuit-output-teacher-free task screen returns seven corrected rejections, none of which places a circuit above a width-matched classical arm, and it establishes no familywise equivalence. Hardware extrapolation is outside scope. We did not measure energy use, so the simulator results do not support an energy-efficiency comparison with classical feature engineering. Several benchmark datasets contain sensitive attributes — Adult predicts income from features including race, sex, age and native country; Churn records customer demographics — and the feature expansions studied here neither remove nor attenuate the biases they carry. Performance on these datasets is not an endorsement for deployment: downstream fairness audits, disparate-impact analysis and, where applicable, mitigation of protected attributes remain necessary.

AI USE STATEMENT

We used generative AI tools to write and validate the experiment runners and statistical guards, to propose and refine hypotheses about circuit design axes, to draft and criticize analysis rules, to re-derive reported numbers from stored arrays during consistency checks, to draft sections of this manuscript, and to translate working documents. We did not use them to generate, augment, or relabel data. Every dataset is a public release; the supplementary code loads the released and recon-

structable inputs, while the omitted primary container and historical regeneration gaps are disclosed in the reproducibility statement and Appendix G. Human-subject and participant-facing uses do not apply to this work. Among the recommended-disclosure categories we also used these tools for code, literature analysis, brainstorming and readability editing. Scripts in the supplementary material produced every figure that plots data; the architecture diagram of Figure 1 is supplied as a separate raster asset. We checked AI-assisted manuscript claims against their cited artifacts and used automated guards where the archive exposes the relevant inputs. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Sercan Ö. Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6679–6687, 2021. doi: 10.1609/aaai.v35i8.16826.

Vasilis Belis, Patrick Odagiu, Michele Grossi, Florentin Reiter, Günther Dissertori, and Sofia Vallecorsa. Guided quantum compression for high dimensional data classification. *Machine Learning: Science and Technology*, 5(3):035010, 2024. doi: 10.1088/2632-2153/ad5fdd.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, B AkashNarayanan, Ali Asadi, et al. PennyLane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018. URL https://arxiv.org/abs/1811.04968.

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

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

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

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Keisuke Fujii and Kohei Nakajima. Harnessing disordered-ensemble quantum dynamics for machine learning. *Physical Review Applied*, 8(2):024030, 2017. doi: 10.1103/PhysRevApplied.8.024030.

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

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pp. 507–520, 2022.

Hans Gundlach, Hrvoje Kukina, Jayson Lynch, and Neil Thompson. Position: Quantum deep learning still needs a quantum leap. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/pdf/de9d722a248b13d4b8a2a2e84782020720afc2b9.pdf.

Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. doi: 10.1038/s41586-019-0980-2.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1):2631, 2021. doi: 10.1038/s41467-021-22539-9.

Naman Jain and Amir Kalev. QuFeX: Quantum feature extraction module for hybrid quantum-classical deep neural networks. *Quantum Science and Technology*, 11(1):015017, 2026. doi: 10.1088/2058-9565/ae24a8.

Siavash Kakavand, Christoph Strohmeyer, and Michael Schlotter. Benchmarking quantum kernel support vector machines against classical baselines on tabular data: A rigorous empirical study with hardware validation. *arXiv preprint arXiv:2604.18837*, 2026. URL https://arxiv.org/abs/2604.18837.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154, 2017.

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

Javier Mancilla and Tomás Tagliani. The fourier wall: Why public tabular datasets refuse quantum advantage, and a certified recipe for where it lives. *arXiv preprint arXiv:2607.15815*, 2026. doi: 10.48550/arXiv.2607.15815. URL https://arxiv.org/abs/2607.15815.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812, 2018. doi: 10.1038/s41467-018-07090-4.

Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018. doi: 10.1103/PhysRevA.98.032309.

Sreeraj Rajindran Nair and Christopher Ferrie. Local surrogates for quantum machine learning. *arXiv preprint arXiv:2506.09425*, 2025. doi: 10.48550/arXiv.2506.09425. URL https://arxiv.org/abs/2506.09425.

Iván Panadero, Yue Ban, Hilario Espinós, Ricardo Puebla, Jorge Casanova, and Erik Torrontegui. Regressions on quantum neural networks at maximal expressivity. *Scientific Reports*, 14(1):31669, 2024. doi: 10.1038/s41598-024-81436-5.

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

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

Arthur Pesah, M Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11(4):041011, 2021. doi: 10.1103/PhysRevX.11.041011.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6638–6648, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007.

Mehrad Sahebi, Alice Barthe, Yudai Suzuki, Zoë Holmes, and Michele Grossi. On dequantization of supervised quantum machine learning via random fourier features. *arXiv preprint arXiv:2505.15902*, 2025. doi: 10.48550/arXiv.2505.15902. URL https://arxiv.org/abs/2505.15902.

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

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

Maria Schuld and Nathan Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019. doi: 10.1103/PhysRevLett.122.040504.

Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *PRX Quantum*, 3(3):030101, 2022. doi: 10.1103/PRXQuantum.3.030101.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021. doi: 10.1103/PhysRevA.103.032430.

Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019. doi: 10.1002/qute.201900070.

Howard Su, Chen-Yu Liu, Samuel Yen-Chi Chen, Kuan-Cheng Chen, and Huan-Hsin Tseng. Scalable quantum machine learning via multi-layer fully-connected variational quantum circuits. *arXiv preprint arXiv:2602.16623*, 2026. URL https://arxiv.org/abs/2602.16623.

Ryan Sweke, Erik Recio-Armengol, Sofiene Jerbi, Elies Gil-Fuster, Bryce Fuller, Jens Eisert, and Johannes Jakob Meyer. Potential and limitations of random Fourier features for dequantizing quantum machine learning. *Quantum*, 9:1640, 2025. doi: 10.22331/q-2025-02-20-1640.

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

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, pp. 682–688, 2000.

C. M. Wilson, J. S. Otterbach, N. Tezak, R. S. Smith, A. M. Polloreno, Peter J. Karalekas, S. Heidel, M. Sohaib Alam, G. E. Crooks, and M. P. da Silva. Quantum kitchen sinks: An algorithm for machine learning on near-term quantum computers. *arXiv preprint arXiv:1806.08321*, 2018. URL https://arxiv.org/abs/1806.08321.

Tianping Zhang, Zheyu Aqa Zhang, Zhiyuan Fan, Haoyan Luo, Fengyuan Liu, Qian Liu, Wei Cao, and Li Jian. OpenFE: Automated feature generation with expert-level performance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41880–41901. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhang23ay.html.

A PIPELINE

Figure 1 introduces the two protocol branches; the detailed pipeline definitions follow below.

For current audit support, Appendices A–C define the protocol-specific pipeline, objectives, and circuit; Appendix E gives the matched controls; Appendix G states artifact scope; and Appendices O–U report cached-map distillation, augmentation, the driver-backed governing five-seed task screen with its historical artifact-only records, the split-encoding hybrid, nomination/transfer, routed-gate counts, and span fits. Retained benchmark context is kept separately in Appendix D and Appendices H–N: benchmark tables, datasets, statistics, downstream models, hyperparameters, design provenance, qubit sensitivity, and circuit diagnostics. Appendix F contains simulator-only noise sensitivities.

The evidential status is not uniform: the benchmark evidence and design grids are exploratory evidence; the controlled-campaign appendices identify their registered or amended status; post-outcome screens are labeled where they appear; and the resource, noise, and circuit diagnostics are simulator-only. This evidence provides no operational rule for choosing the circuit over Poly2, learned RFF, a tuned MLP, or raw-feature boosting.

Figure 1 shows the whole assembly; the rest of this appendix gives the parts. Its register is drawn at $q{=}4$, matching the convention of Figure 4; that register would give $4q{-}1{=}15$ features, whereas the unlabeled 23 in its measurement box is the primary $q{=}6$ count. Augmentation classification uses one identity-initialized circuit and $4q{-}1$ features. Appendix B writes out the preprocessing path, the trainable compression to q qubits and the augmentation training objective; Appendix C draws the circuit block at gate level and counts the $4q{-}1$ features one circuit emits. In the regression augmentation branch, three independently fitted full transforms give a width-$3(4q{-}1)$ block (23 and 69 only at $q{=}6$). In each controlled split and seed, every named arm deterministically reconstructs the same frozen training-only MI–quantile–PCA–affine front end; equality of split, compression, and coordinate hashes is required. MI reads y_{train}, so this front end is not unsupervised or target-blind; it is fixed before circuit training and is not co-trained with the circuit. The affine map sends the training PCA range to $[-1, 1]$ without clipping held-out coordinates, and held-out overshoot is recorded. Each controlled arm emits one width-$(4q{-}1)$ map. These are distinct measured objects, not two views of one shared fit.

B PROTOCOL-SPECIFIC COMPRESSION AND TRAINING OBJECTIVES

Section 3 names the augmentation's trainable compression layer and hybrid training loss; both definitions follow. Controlled arms instead freeze their MI–quantile–PCA–affine front end and use the CKA objective specified below.

Trainable compression layer. A central challenge in applying quantum circuits to real-world tabular data is that the input width n can exceed the number of qubits. Primary cells use $q{=}6$, while dedicated sensitivity and cost sweeps vary q. MI preselection retains $\tilde{n} = \min(n, 64)$ columns. If MI is needed, it may be estimated on a seeded subsample of at most 5,000 training rows; after ranking, the retained columns are returned to their original column order. A QuantileTransformer is then fitted on those training columns with $n_{\text{quantiles}} = \min(1000, N_{\text{train}})$ and its uniform outputs are scaled to $[0.05, \pi - 0.05]$. Only after these steps is the supervised compression fitted. For a column vector $\mathbf{x} \in \mathbb{R}^{\tilde{n}}$, the implemented map is

$$\mathbf{z} = \begin{cases} \tanh\!\left(\mathbf{W}\mathbf{x}/\sqrt{\tilde{n}}\right), & \tilde{n}/q \geq 10, \\ \cos(\mathbf{W}\mathbf{x} + \mathbf{b}), & \tilde{n}/q < 10, \end{cases} \quad \mathbf{W} \in \mathbb{R}^{q \times \tilde{n}}, \quad \mathbf{b} \in \mathbb{R}^{q}, \tag{1}$$

where $\rho_{\text{enc}} = \tilde{n}/q$ drives both this branch and the augmentation loss gate below. The implementation allocates $\tilde{n}q + q$ compression slots, but the high-ratio branch does not read the q bias slots; the low-ratio branch reads all of them. The activation choice was not isolated from the rest of the compound transform. Every component above is fitted on the training partition and then applied unchanged to held-out rows.

The PCA-labeled augmentation member retains child seed $s{+}2000$ and first completes the identity-initialized fit. It reuses that member's fitted selector and QuantileTransformer, fits

PCA at seed $s{+}2500$ on at most 2,000 scaled training rows, overwrites the compression, resets the compression bias and the encoding, wide, and deep parameters, and retrains on at most $\min(\max(700, 5\sqrt{N}), 5000)$ training rows. The identity and random members retain child seeds s and $s{+}1000$.

Augmentation training objective. After fitting MI selection and the quantile transform, each augmentation member jointly optimizes its compression layer and circuit parameters using a hybrid loss. Let $\mathbf{q}(\mathbf{x})$ denote the $(4q{-}1)$-dimensional quantum feature vector. The loss combines two terms:

Target-alignment loss. A mean squared error term that steers the circuit readout toward the supervised target:

$$\mathcal{L}_{\text{align}} = \frac{1}{N}\sum_{i=1}^{N}\left(\overline{\langle Z\rangle}(\mathbf{x}_i) - \tilde{y}_i\right)^2, \tag{2}$$

Here $\overline{\langle Z\rangle}(\mathbf{x}) = q^{-1}\sum_{j=1}^{q} q_j^Z(\mathbf{x})$ is the mean of the q single-qubit Pauli-Z expectations contained in $\mathbf{q}(\mathbf{x})$. The training target $\tilde{y}$ is standardized to zero mean and unit variance for regression. For classification it is a binary class indicator, taken one-versus-rest against the first class when more than two classes are present. Only this scalar summary enters the training term. At inference the full $(4q{-}1)$-dimensional $\mathbf{q}(\mathbf{x})$ is retained; the augmentation protocol concatenates the three readouts with the original input, whereas a downstream head scores each controlled arm's single readout map without original-feature concatenation.

Correlation objective. For centered batch vectors, define the epsilon-stabilized correlation $r_{uv} = \langle u-\bar{u}, v-\bar{v}\rangle / [\sqrt{\|u-\bar{u}\|_2^2+\epsilon}\sqrt{\|v-\bar{v}\|_2^2+\epsilon}]$, with $\epsilon = 10^{-8}$. The implemented term is

$$\mathcal{L}_{\text{corr-div}} = 1 - \frac{1}{m}\sum_{k=1}^{m} r_{q_k y}^2 + \lambda \frac{1}{m(m-1)}\sum_{j\neq k} r_{q_j q_k}^2. \tag{3}$$

Here $m{=}4q{-}1$ (23 at $q{=}6$), and the loss is applied separately to each full transform in the three-member augmentation stack. We set $\lambda{=}0.3$ after a 10-dataset, 5-seed sweep over $\{0, 0.01, 0.1, 0.3, 0.5\}$; because that cohort overlaps the evaluation benchmark, this defines the evaluated regime rather than an independently confirmed optimum. The objective trains the compound transform and does not isolate a contribution from its circuit component.

The two losses are combined with a sigmoid-gated mixing coefficient that adapts to the compression ratio:

$$\mathcal{L} = \alpha\cdot\mathcal{L}_{\text{align}} + (1-\alpha)\cdot\mathcal{L}_{\text{corr-div}}, \quad \alpha = \text{sigmoid}(0.5\cdot(\rho_{\text{enc}} - 8)). \tag{4}$$

When $\rho_{\text{enc}} \gg 8$ the target-alignment term receives more weight; when $\rho_{\text{enc}} \ll 8$ the correlation term does. Both gates act on ρ_{enc}: the activation threshold $\rho_{\text{thr}}{=}10$ and the sigmoid center 8 are fixed across all 37 datasets, and the offset separates them.

The augmentation compression and circuit parameters are optimized jointly using PennyLane 0.38.0 backpropagation on the `default.qubit` analytic-statevector simulator (Bergholm et al., 2018). Adam starts at learning rate 0.01, multiplied by 0.95 every 10 steps, with batch size 64 and at most 100 steps. On batch wraparound the training rows are reshuffled and batching restarts. Training stops after 20 consecutive steps without an improvement greater than 10^{-4}; the parameters at stopping are retained, not a best-loss checkpoint.

Controlled circuit objective. Controlled arms do not use Equation (4). With centered circuit readouts Q_c and centered label factor Y_c, they maximize

$$\text{CKA}(Q, Y) = \frac{\|Q_c^\top Y_c\|_F^2}{\|Q_c^\top Q_c\|_F \, \|Y_c^\top Y_c\|_F}. \tag{5}$$

Regression uses a standardized scalar label factor; classification uses a class-balanced one-hot factor. Full-batch Adam runs at learning rate 0.01 for 300 steps. A fixed, disjoint 20% inner holdout receives no gradients; Equation (5) is evaluated on it every 10 steps and the best holdout checkpoint is retained. The deterministic fit and selection subsets are capped at 1,024 and 2,048 rows, respectively. Cap subsampling uses a NumPy random state keyed by the cell seed and the first eight

hexadecimal digits of the training-coordinate hash modulo 2^{31}, so the row sets do not depend on the controlled-arm name.

Two diagnostics of the fitted map follow. Figure 2 displays the trained compression weights themselves across datasets and stack members, and Figure 3 the training dynamics of the readout those weights feed. Both describe one fitted configuration and enter no contrast family.

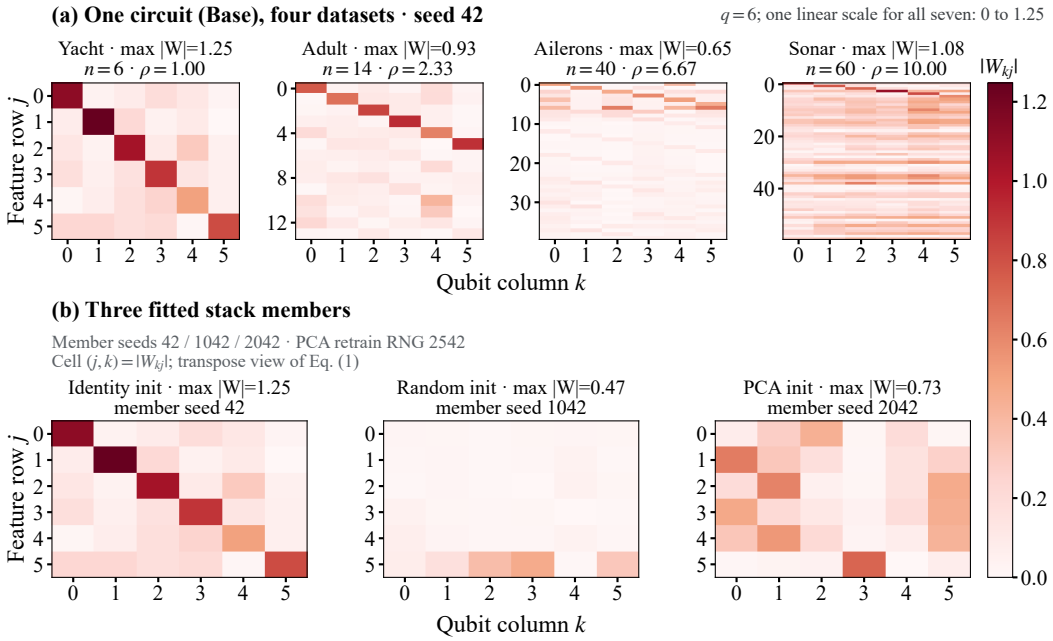

Figure 2: Absolute trained compression weights from Equation (1), displayed with selected-feature row j and qubit column k: cell (j, k) is $|W_{kj}|$, the transpose view of the equation's $q \times \tilde{n}$ matrix $\mathbf{W}$. All seven panels use one linear light-rose–to–burgundy scale over $[0, 1.25]$, and the shared colorbar reports $|W_{kj}|$; each title also reports its panel maximum. (a) The augmentation 1W+2D single circuit at seed 42 on four selected datasets spanning $\rho_{\text{enc}} = \tilde{n}/q = n/q$ from 1.00 to 10.00; raw and selected widths coincide in these panels. The embedded panel label "Base" denotes this augmentation 1W+2D single circuit, not controlled C0. Because dataset identity and raw feature semantics vary jointly with ρ_{enc}, the row is descriptive rather than a controlled ρ_{enc} sweep. (b) The three fitted members of one regression stack on Yacht, using identity, random, and PCA initialization with member seeds 42, 1042, and 2042; the PCA overwrite/retrain uses RNG seed 2542. Initialization and member seed therefore vary jointly, so the panels characterize the fitted members rather than isolate an initialization effect. The identity member and the Yacht single circuit in (a) return bit-identical weights. The random member has maximum $|W_{kj}| = 0.466$, 37% of the shared maximum.

C Quantum Circuit Diagram

Figure 4 illustrates C0 at $q = 4$; its four stages are as follows.

Stage 1: Encoding. Each qubit j receives the compressed input z_j via two parameterized rotation gates, $\mathrm{R}_X(w_j^{(x)} z_j + b_j^{(x)})$ and $\mathrm{R}_Y(w_j^{(y)} z_j + b_j^{(y)})$, where $w_j^{(x)}, w_j^{(y)}, b_j^{(x)}, b_j^{(y)}$ are per-qubit encoding scalars, distinct from the weights of the trainable compression front end of Section 3. This is the *only* point where classical data enters the circuit; all subsequent operations are data-independent and do not re-upload input data. For $q = 6$ qubits, the encoding layer contributes $4q = 24$ trainable parameters.

Stage 2: Wide layer. A general three-Euler-angle rotation $\mathrm{Rot}(\theta, \phi, \lambda)$ is applied to each qubit, followed by a nearest-neighbor CNOT chain ($j \to j{+}1$ for $j = 1, \ldots, q{-}1$). The chain topology,

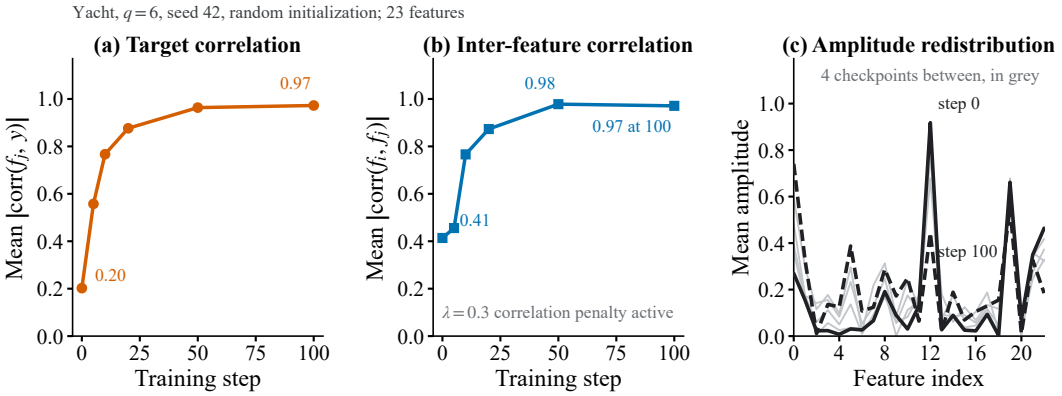

Figure 3: Training dynamics of the $(4q-1)$-dimensional augmentation readout (23 dimensions here at q=6) for the fitted random-initialized member on Yacht at seed 42. (a) Mean |corr| between each readout feature and the target. (b) Mean |corr| between feature pairs over the same run, with the λ=0.3 pairwise-correlation penalty of Equation (3) active throughout. The increase in inter-feature correlation is one diagnostic trajectory and does not identify its cause. (c) Mean amplitude per feature at the first and last checkpoint, the four intermediate checkpoints in grey. Section 4 measures how well classical students reproduce this readout.

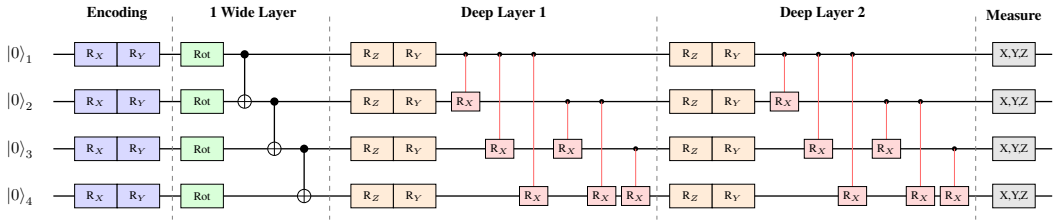

Figure 4: The C0 1W+2D circuit at gate level, drawn for $q = 4$. Primary benchmark cells use $q = 6$, while dedicated sensitivity and cost sweeps vary q. The adjacent ZZ correlations that complete the $3q + (q-1)$ readout are measured but not drawn.

highlighted with green-shaded Rot gates in Figure 4, produces $q-1$ entangling gates — three in the q=4 drawing, five at the primary q=6 — creating local correlations between adjacent qubits with minimal circuit depth. This layer contributes $3q$ parameters, 18 at q=6.

Stage 3: Deep layers ($\times$2)**.** Each deep layer applies per-qubit R_Z and R_Y rotations followed by the all-to-all CR_X entanglement drawn in red in Figure 4. For q qubits, the all-to-all topology creates $\binom{q}{2}$ entangling gates per deep layer, 15 at q=6, enabling rich inter-qubit correlations that complement the nearest-neighbor structure of the wide layer. The $\binom{q}{2}$ entangling gates in a deep layer share a single angle per control qubit, and the last qubit controls none, so each deep layer contributes $2q + (q-1)$ parameters, 17 at q=6, totaling 34 for both.

Stage 4: Multi-basis measurement. Each qubit is measured in all three Pauli bases (σ_X, σ_Y, σ_Z), yielding $3q$ single-qubit expectation values. Additionally, adjacent ZZ correlations $\langle \sigma_Z^{(j)} \otimes \sigma_Z^{(j+1)} \rangle$ are measured for $q-1$ pairs, omitted from Figure 4 for visual clarity. This produces $3q + (q-1) = 4q-1$ quantum features per circuit evaluation, which is 23 at q=6. We restrict two-qubit measurements to adjacent ZZ pairs, which holds the readout width linear in q; we did not benchmark an all-pair readout.

Design rationale. The wide layer supplies a nearest-neighbor CNOT chain and the deep layers use all-to-all controlled-R_X gates. At q=6, the augmentation 1W+2D implementation allocates 90 circuit slots but gates read 76 of them: 24 encoding, 18 wide-layer, and 34 deep-layer parameters.

The other 14 allocated slots are not read by any gate. No functional- or Jacobian-rank count is established.

Named controlled variants. The controlled specification allocates only gate-acting parameters. C0 is the controlled name for the 1W+2D ansatz and has one encoding block, one wide layer, and two deep layers, with one entangling angle per control qubit. C-edge retains the emitted gates but assigns a separate angle to every deep-layer edge. C-GraphRZZ changes the data encoding and reachable function class as well as gate family and topology. For each split or fold, it selects the top $q-1$ pairs by absolute Spearman dependence on compressed training coordinates, breaks ties by channel indices, and freezes the rebuilt graph for evaluation; edge (i, j) applies $\mathrm{IsingZZ}(\theta_{ij} z_i z_j)$. Controlled C-reupload applies three repeated wide–deep stacks with three block-specific $q \times 4$ encoding arrays (separate RX/RY slopes and offsets), while tying the wide, deep, and entangler tensors across all stacks. Their gate-acting counts at q=6 are 76, 96, 76, and $3(4q) + 3q + 2[2q + (q-1)] = 124$, respectively, while every arm retains the same $(4q-1)$-dimensional readout (23 at q=6). These named architectural contrasts are width-matched, but not parameter- or compute-matched, and hence do not isolate one causal mechanism. The augmentation implementation's re-upload mode is a different construction: it inserts the same learned encoding parameters again before selected layers rather than fitting block-specific encoding arrays.

D BENCHMARK RESULT TABLES

The two tables report the retained benchmark record, including all score columns and the classification results. Their cells are transcribed from the stored result artifacts and are not recomputed for this appendix.

Table 2 is that record at full width — the whole k=25 field on one page, with the seven ablation arms printed as rows rather than left to compete without appearing — and it is here because two things elsewhere in the paper cannot be checked without it: the in-sample architecture settings of Appendix L were read off this field, and every score in it comes from the best-of-k protocol that Appendix J sets against the selection-free tables it treats as primary.

In this benchmark record, QFX denotes the evaluated quantum feature expansion. Read down the *Sig.* column, that field separates very little. Eight arms exceed the critical difference against QFX, and every one of them sits at overall rank (OVR) 14.46 or worse: the three deep tabular baselines, the two end-to-end quantum predictors, CoVar, ZZFeatureMap and QFX-QO, which is QFX with the original-feature concatenation removed so that the readout sees the $4q-1$ circuit outputs alone; its layer counts, circuit count and re-uploading setting are those of QFX. QFX is separated from no classical feature map — not from Poly2 at 8.01, nor from PCA, Original, RFF (Rahimi & Recht, 2007), Nyström (Williams & Seeger, 2000) or OpenFE — and from none of PVQNN, QRC, QKS or PQK. A rank of 7.81 at the head of 25 methods is therefore a position in an undifferentiated field, not a separation from one.

Figure 5 plots each dataset's rank from this field against its samples-per-feature ratio. The N/n rule drawn there was read off these same 37 datasets after the ranks existed, so it locates a pattern and is not a selection rule.

Table 3 is the narrower of the two and the one Section 5 puts back under measurement: its two regression score columns are the cells that support the circuit's lead over every classical feature map in its pool under fixed penalized linear readouts. On regression every classical Ridge mean is below 66 or Poly2 collapses, and every classical Lasso mean is at most 12.2 against the 61.94 of QFX; no sole cause for the Poly2 collapse is established. That 12.2 ceiling belongs to this pool rather than to the readout — Section 5 adds four transformers the pool did not contain, and the circuit then ranks fourth of ten with 49.17 pp of the Lasso margin attributable to the pool. Two changes separate these two views and only one of them is the pool. The 61.94 above is this unsealed benchmark grid; on the sealed remeasurement the circuit's own Lasso T0 score is 57.85 (Table 9), and it is against 57.85, not 61.94, that exactly three of the four added arms stand higher and produce the fourth-of-ten rank. The 4.09 pp the circuit itself loses under sealing is not attributable to pool expansion and is not reconciled here.

Table 2: Overall performance and rank over 37 datasets (17 regression, 20 classification) for the whole k=25 field; the final block is the benchmark ablation study. This is the retained benchmark grid, not the sealed remeasurement of Section 5: its splits are unsealed and every score is a maximum over downstream estimators taken on the test split, with no validation partition, so the numbers are optimistic as absolute estimates and are not the primary view. Appendix J states that boundary and treats the fixed-estimator tables as primary where the two disagree; the five end-to-end predictors — FC-VQC and QML-Direct in the block so named, plus FT-Transformer, TabNet and Tuned MLP, none of which takes a downstream model — receive none of that maximization premium, which averages $+1.57$ pp on regression and $+0.97$ pp on classification. Scores are $\times 100$, the per-dataset best across downstream models averaged over 5 seeds, $\pm$ the mean per-dataset SD; ranks are per-dataset among the 25 methods, averaged within task for REG and CLS and over all 37 for OVR. Dashes are entries the retained record does not carry. ΔOVR is a row's OVR minus the 7.81 of QFX, and *Sig.* reads *yes* where $|\Delta\text{OVR}|$ exceeds the Nemenyi critical difference CD=6.26 at α=0.05 (Demšar, 2006); #Wins credits every method in a tie; four datasets are tied, so the column sums to 41 rather than 37. QFX takes three, and all three are ties with Base, so it wins no dataset outright. **Bold** marks the best value in a column and underline the second best, at the printed precision with ties sharing a mark; in rank columns best is lowest, and later tables follow this convention. Appendix I states the FT-Transformer and TabNet caveats.

Method	Overall	Regression	Classification	Avg. Rank			vs. QFX		#Wins
				REG	CLS	OVR	ΔOVR	Sig.	
Classical Baselines									
Original	81.23±1.51	74.73±1.69	86.76±1.35	12.76	9.22	10.85	+3.04	no	1
PCA	81.79±1.42	75.45±1.47	**87.17±1.37**	10.88	**7.38**	8.99	+1.18	no	1
Poly2	**81.98±1.38**	76.22±1.50	86.88±1.28	6.71	9.12	8.01	+0.20	no	**7**
RFF	81.69±1.52	75.52±1.64	86.94±1.42	13.35	10.85	12.00	+4.19	no	2
Nyström	81.73±1.56	75.61±1.72	86.93±1.43	9.41	10.62	10.07	+2.26	no	2
OpenFE	81.33±2.13	**76.36±2.08**	85.56±2.17	8.24	12.47	10.53	+2.72	no	3
Deep Tabular Baselines									
FT-Transformer	73.02±3.16	66.16±4.44	78.85±2.08	15.18	21.30	18.49	+10.68	yes	3
TabNet	32.32±100.46	−24.07±215.36	80.25±2.79	17.65	21.75	19.86	+12.05	yes	1
Tuned MLP	77.78±6.26	71.24±11.11	83.35±2.13	10.29	18.98	14.99	+7.18	yes	4
Quantum Feature Maps									
PVQNN	81.15±1.60	74.22±1.72	87.04±1.50	14.47	8.32	11.15	+3.34	no	1
QRC	81.07±1.54	74.35±1.70	86.78±1.41	15.59	8.90	11.97	+4.16	no	0
QKS	80.87±1.43	73.90±1.74	86.79±1.17	16.00	11.82	13.74	+5.93	no	1
PQK	80.41±1.44	73.43±1.62	86.34±1.29	16.88	11.28	13.85	+6.04	no	3
CoVar	80.56±1.49	73.66±1.69	86.43±1.32	16.88	12.50	14.51	+6.70	yes	0
ZZFeatureMap	80.42±1.51	73.49±1.75	86.31±1.32	17.35	12.00	14.46	+6.65	yes	0
End-to-End Quantum									
FC-VQC	56.40±8.44	41.49±13.40	69.08±4.23	23.71	20.70	22.08	+14.27	yes	2
QML-Direct	45.07±8.33	26.78±14.00	60.62±3.51	24.41	24.07	24.23	+16.42	yes	0
Ours									
QFX	81.97±1.47	76.24±1.65	86.84±1.31	**6.26**	9.12	**7.81**	0	ref.	3
Benchmark ablation arms									
Base	81.68	75.60	86.84	10.06	9.12	9.55	+1.74	no	4
Deep-Only	81.53	75.45	86.69	—	—	10.82	+3.01	no	0
Wide-Only	81.49	75.53	86.56	—	—	12.09	+4.28	no	0
All-Stack	81.80	76.24	86.53	**6.26**	11.60	9.15	+1.34	no	1
Deeper	81.84	76.24	86.60	—	—	8.50	+0.69	no	0
+Reupload	81.35	75.16	86.61	—	—	9.73	+1.92	no	1
QFX-QO	72.99	72.45	73.45	—	—	17.55	+9.74	yes	1

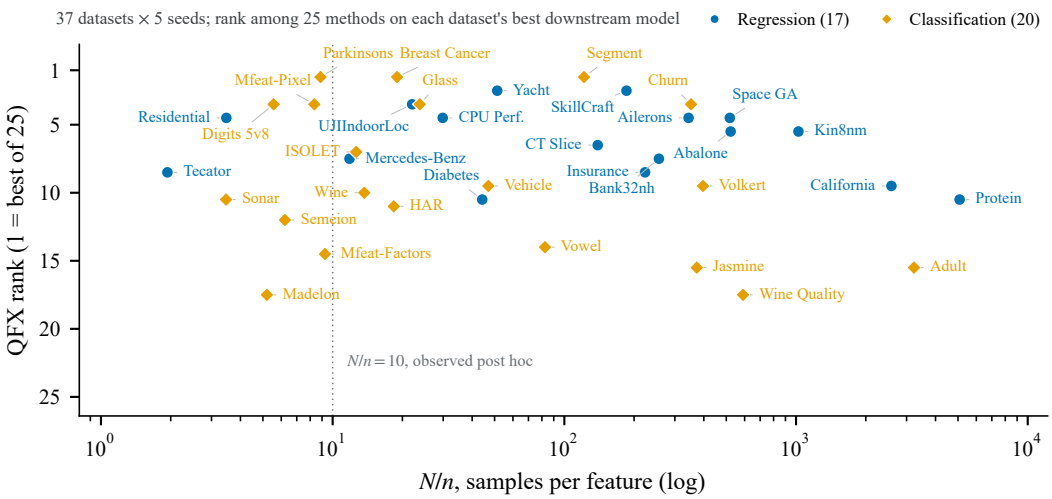

Figure 5: Per-dataset rank of QFX among the k=25 methods of Table 2, against samples per feature. Each point is one dataset's rank on its own best downstream model, averaged over 5 seeds. Color and marker separate the 17 regression from the 20 classification datasets. The dotted line at $N/n = 10$ was observed after these ranks were computed rather than registered before them; it locates the pattern and is not a selection rule.

Table 3: Ridge and Lasso downstream performance, $\times 100$: 5-seed per-dataset means over the 17 regression or 20 classification datasets, with per-task ranks across the 12 methods shown. PQK is omitted. This is the retained benchmark grid, not the sealed remeasurement used for the paired contrast in Section 5. The retained classification *Lasso* score and rank columns have incomplete, method-dependent coverage (Appendix G); their cells are incomparable and receive no bold or underline emphasis. Classification headings denote penalized logistic regressions (Table 7). Section 5 re-measures the Lasso margin printed here.

	Regression (R^2)		Classification (Acc.)		Ridge Rank (↓)		Lasso Rank (↓)	
Method	Ridge	Lasso	Ridge	Lasso	REG	CLS	REG	CLS
Classical Baselines								
Original	8.14	3.08	81.83	80.48	9.29	7.65	8.76	6.53
PCA	18.52	12.18	82.16	82.69	9.03	5.60	8.06	5.78
Poly2	-2.27×10^6	-2.20×10^6	82.55	84.29	6.12	8.30	<u>3.94</u>	5.43
RFF	56.19	5.96	83.55	83.48	4.12	**4.62**	6.79	5.00
Nyström	<u>65.96</u>	8.14	<u>83.57</u>	83.27	<u>3.06</u>	**4.62**	5.85	4.27
OpenFE	40.73	6.08	79.26	80.94	6.03	9.20	5.32	6.15
Quantum Feature Methods								
PVQNN	48.77	28.62	**83.83**	83.27	6.29	5.03	5.91	4.29
QRC	49.25	13.51	82.48	81.43	6.53	6.20	7.03	7.17
QKS	54.98	<u>36.10</u>	83.17	82.41	7.59	6.15	6.97	6.04
ZZFeatureMap	−8.16	4.18	82.12	80.10	8.47	7.97	8.44	9.08
CoVar	42.80	32.12	81.89	80.50	8.65	7.97	7.76	8.92
QFX	**72.15**	**61.94**	83.20	82.39	**2.82**	<u>4.67</u>	**3.15**	5.08

Figure 6 sets the same comparator pool against the seconds each arm spends to reach its regression Ridge score. Fit seconds are wall-clock under one recorded environment and carry no hardware or deployment claim.

The classification Lasso column of Table 3 averages an uneven cohort — 17 datasets for Original down to 12 for QFX and every other quantum arm (Appendix G) — so score comparisons down that column are not like-for-like and we make none. OpenFE appears in the benchmark record; the sealed measurement excludes it as a comparator — its transform is not row-local (on Insurance, 10 of 29 generated columns depend on other test rows) and its sealed cohort is 4 of 17 — while

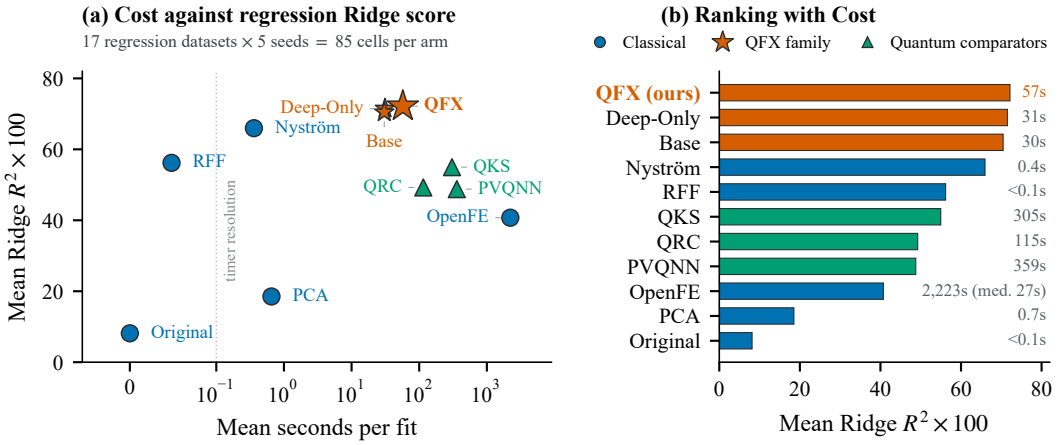

Figure 6: Cost against regression Ridge score for the benchmark comparator pool. (a) Mean seconds per fit against mean Ridge $R^2 \times 100$, both over the 17 regression datasets at 5 seeds; the dotted line marks the timer resolution, below which the sub-second arms are not separated. (b) The same scores ordered, each bar carrying that arm's mean fit time. The eleven arms shown are drawn from the same grid as Table 3 but are not that table's twelve: nine are common to both, this figure adds Base and Deep-Only, and it omits Poly2, ZZFeatureMap and CoVar. All eleven timings are recomputed under one convention, the per-(dataset, seed) mean over those 17 datasets, and are wall-clock under 7–10 concurrent workers: comparable with each other, not machine-normalized. OpenFE's mean is set by a few datasets and its median is 27 s. The label "QFX (ours)" in panel (b) is the benchmark figure's original legend; the audited family is the authors' prior design (Section 3), and no method claim attaches to the label.

its benchmark scores are kept because the pre-registered Holm family may not shrink after the fact (Section 5).

ewpage removed 2026-09-25 (C-59): Appendix E no longer forces a page break; the half-empty page 22 it produced is gone

E Matched optimizer and parameter controls under Lasso and Ridge

Matched Lasso full-joint-optimizer control. Across 13 datasets and five seeds, both arms were generated side by side in the same cell process and environment. They use q=6, the three-member identity/random/PCA construction, 1W+2D circuits, $[x, q(x)]$, T1 fold-local/outer-training `StandardScaler`, and Lasso α=0.01. The intervention switches the complete supervised joint optimizer bundle—compression weight and bias, encoding, wide, and deep parameters—against a zero-step control; the component ablation below partitions the bundle by trainable group. All 65/65 cells succeeded; all 260 optimizer-entry equality comparisons pass, covering 185 distinct hashes. The optimizer-on arm records 19,346 training steps versus zero in the control arm; these counts are not timing evidence. After averaging seeds within dataset, optimizer-on and zero-step scores are 66.693014 and 65.556021 in R^2×100: the optimizer-on-minus-zero-step mean is +1.136993 pp, the median is +0.638586 pp, the record is 11–1–1, and the dataset bootstrap 95% interval for the mean is [+0.360042, +2.081086] pp. This dataset-bootstrap interval quantifies reweighting sensitivity of the amended 13-dataset decision cohort, not uncertainty for a task superpopulation. The interval excludes zero, so the full optimizer-on bundle helps directionally in this matched comparison; because both summaries fail to exceed the strict 2.0 pp E7 convention, it supports no practical-threshold claim. This control does not isolate a C-reupload mechanism.

Component-wise optimizer ablation. Run `e7_paired_component_ablation_20260818` generates five arms side by side in each of the same 65 cells: the full bundle, an encoding-plus-gate-angle arm (`params_encoding`, `params_wide`, `params_deep`), a gate-angle-only

arm (params_wide, params_deep), a compression-only arm (params_compression, params_compression_bias), and the zero-step control; groups outside an arm's set stay frozen at the shared initialization. All four optimizer-entry full-parameter hashes match across every arm before any step (65 cells $\times$ 4 entries $\times$ 4 reference comparisons, 1,040 equalities), every frozen group's post-fit hash equals its entry hash, and the full-bundle and zero-step arms reproduce the matched control above exactly, cell by cell. After averaging seeds within dataset, the prespecified primary contrast, gate-angle-only minus zero-step, is $+0.050088$ pp mean, 0.000000 median, 6–6–1, with dataset-bootstrap 95% interval $[-0.260652, +0.383218]$ pp (100,000 resamples, seed 20260818). Compression-only minus zero-step is $+1.026759$ pp mean, $+0.775576$ median, 10–2–1, interval $[+0.408344, +1.880804]$ pp; encoding-plus-gate-angles minus zero-step is -0.016222 pp mean, $+0.075262$ median, 9–3–1, interval $[-0.784233, +0.547183]$ pp; and full bundle minus encoding-plus-gate-angles is $+1.153215$ pp mean, $+0.143566$ median, 9–3–1, interval $[+0.115983, +2.488543]$ pp. Early stopping makes the arms not step-matched: executed step totals are 19,346 (full), 15,571 (encoding-plus-angles), 15,276 (gate angles), 16,301 (compression), and 0 (zero-step control); the step-matched sensitivity is reported below, and these counts are not timing evidence. Within this roster and protocol, compression-only training recovers most of the matched control's bundle gain, and no isolated gate-angle gain is detected under this non-step-matched protocol. The frozen-front-end Ridge protocol below differs in readout and freeze point: it freezes the compression layer for both arms and compares trained against fixed random angles, whereas this ablation trains group subsets from the shared initialization inside the Lasso pipeline. The ledger SHA-256 is fc0fbd80ee61a0e1d0e29507023f5320c0b635bf82bd58b80a76b1f85b190a59; the analysis lock SHA-256 is e7115041114fc7780648edf79055282c4c56504a940004a0ba37d6bf3077f07a.

Step-matched component ablation. The run e7_step_matched_component_ablation_20260921 repeats the five-arm ablation on the same 65 cells with the step confound removed. Regime M fixes, at each of the four optimizer entries, every partial arm's step count to the count the full bundle executed in that cell (19,346 steps for every trained arm); regime F fixes every arm, the full bundle included, at the 100-step cap (26,000). All nine arms enter every entry from identical parameter hashes, frozen groups do not move, executed counts are read from the optimizer boundary rather than from the configured target, and the trained and zero_step arms reproduce the locked control cell by cell (largest deviation 3.8×10^{-6} pp, a cross-machine rounding effect recorded in the lock's amendment). After averaging seeds within dataset, gate-angle-only minus zero-step is -0.101 pp mean, $+0.025$ median, 7–5–1, dataset-bootstrap 95% interval $[-0.513, +0.263]$ pp under M, and $+0.002/0.000$ pp, 7–5–1, $[-0.320, +0.320]$ under F; compression-only minus zero-step is $+1.291/+0.821$ pp, 11–1–1, $[+0.509, +2.239]$ (M) and $+1.371/+0.680$ pp, 12–0–1 (F); the matched gate-angle arm sits 0.151 pp below its non-matched sibling ($[-0.319, -0.021]$), and the full bundle at the cap is indistinguishable from the early-stopped bundle (-0.002 pp). Removing the step difference therefore creates no gate-angle effect; the attribution of the bundle to the classical compression layer stands under both protocols. This is a post-hoc paired sensitivity on a reused cohort, not untouched confirmation, and carries no timing claim. The ledger SHA-256 is f6a07ad745562eacd284ea6413a054e828e1c9cafc9bbf0a824652706bb9c5d4; the lock SHA-256 is efec6577800df781d0d43e74d09906204e25b5117baa280b596fef46219aaed2.

Separate Ridge control. Section 5 measures a trained circuit against an untrained quantum map and finds a gap. The single-factor attribution protocol below instead compares the same circuit at the same register size on the same compressed inputs, with only the parameter tensor varying, and finds no separation. Its values are transcribed from the retained result artifacts.

The Ridge and QKS contrasts are not the same measurement. Under the separate Ridge protocol below, both arms consume one compression layer that is fitted once per split and then frozen, both run at the q=6 this paper fixes, and the untrained arm is that same circuit carrying fixed random parameters. The comparator in Section 5 is the benchmark pool's QKS arm, whose circuit family, register sizing and Z-basis-only readout are specified in Appendix K and which reaches the data without the trainable front end of Section 3. A gap there separates two pipelines; a gap in this Ridge comparison would separate two settings of one circuit's parameters, and that Ridge gap does not emerge. The same-run paired Lasso control above is a different readout protocol and does show a

Table 4: Frozen-front-end attribution over all 17 regression datasets, 5 seeds, q=6. Relative to the arm-neutral protocol, the compression layer is fitted once per split on QFX and then frozen and shared across arms, the downstream model is Ridge alone, and p-values are Holm-corrected—Holm_5 within a subset and Holm_{10} jointly. The nonlinear subset contains the 9 datasets fixed by the pre-specified rule in the text. Gaps are in percentage points, positive favoring QFX; exact Wilcoxon p-values are dyadic and rounded half up. * marks a Holm-corrected $p < 0.05$ and is independent of the global bold/underline rank-emphasis convention.

	All 17				Nonlinear 9			
Comparison	Mean	W–L	Holm_5	Holm_{10}	Mean	W–L	Holm_5	Holm_{10}
vs. bypass	+2.61	15–2	.0067*	.0134*	+4.77	9–0	.0195*	.0352*
vs. MLP head	+1.24	14–3	.0374*	.0654	+2.28	8–1	.0313*	.0625
vs. RFF head	+0.81	15–2	.0386*	.0772	+1.67	8–1	.0820	.1367
vs. random circuit	+1.00	13–4	.0895	.1791	+2.04	8–1	.1484	.2227
vs. Poly2 head	−0.10	7–10	.8536	1.0000	−0.22	3–6	.6523	1.0000

directional full-joint-optimizer-on versus zero-step gain below the 2.0 pp convention; its component ablation locates that gain in the classical compression layer, consistent with the null angle separation here.

Protocol. This frozen-front-end protocol is the stricter of the two retained attribution protocols. Six arms are scored: a compression-only bypass that feeds the compressed vector straight to the downstream model, a fixed random circuit, a dimension-matched Poly2 head, a dimension-matched RFF head, a trained `tanh`-MLP head, and the full trained circuit of QFX. The compression layer is fitted once per split and then frozen and shared across all arms, the downstream model is fixed Ridge alone, and p-values are Holm-corrected over the pre-specified comparisons. One caveat is structural and we state it plainly: the shared front end is fitted on QFX, not arm-neutrally, so this protocol is stricter about compression variance but is not itself a neutral arbiter between arms.

Table 4 prints the whole protocol rather than only the contrasts that pass, so the trained-versus-fixed-random comparison can be read beside the four surrounding controls. This protocol directly reaches 13 of the 17 regression datasets because it reads only one of two dataset stores. The separately stored Ailerons, Bank32nh, Residential, and UJIIndoorLoc datasets were evaluated under the identical protocol (Appendix G), and the table reports all 17. On the 13 directly accessible datasets, the bypass gap is $+3.25$ pp on the mean and $+1.80$ pp on the median, with a SkillCraft collapse separating the statistics; full-17 medians are not retained. The Holm_{10} column corrects over the joint ten-test family rather than the five contrasts within a subset. The 72.15 and 65.96 in Table 3 are unclipped, whereas clipping negative R^2 at zero raises both, so the two conventions must not be compared directly.

Extension to all 17 regression datasets. Under the identical protocol, the MLP and RFF contrasts newly pass Holm_5 at $n = 17$ (.0374 and .0386), while bypass—already significant under Holm_5 at $n = 13$—also passes the broader Holm_{10} family at .0134. No sign reverses.

Why one effect size shrinks while tests strengthen. For the bypass comparison specifically, the four added datasets contribute a mean $+0.56$ pp and point the same way in all four cases. Across all five comparators their mean contribution is $+0.24$ pp, and Poly2 and Random each have a 3W/1L record. Small, mostly consistent additions can reduce a mean while strengthening a signed-rank test; no stronger common-direction claim is made.

The nonlinear subset is fixed in advance, not chosen afterwards. Membership follows a rule registered before this run: a dataset is nonlinear if $\max(\text{Original-XGB}, \text{Original-Tree}) - \text{Original-Ridge} > 5$ pp. Applying it to the four new datasets gives 1.09, 0.86, 3.28, and 4.83 pp for Ailerons, Bank32nh, Residential, and UJIIndoorLoc, all below the threshold, though UJIIndoorLoc misses it by only 0.17 pp. The subset therefore stays at 9 datasets. No alternative subset was considered after seeing the results.

What this does not establish. Two negative results persist across both cohorts and we do not soften them. The dimension-matched Poly2 head does not separate from the trained circuit—-0.10 pp over all 17 datasets (7W/10L, $p = .854$) and -0.22 pp on the nonlinear subset (3W/6L, $p = .652$)—so we cannot claim that QFX significantly beats every matched classical nonlinear head. The trained circuit also fails to separate significantly from a fixed random circuit under this Ridge protocol ($+0.998$ pp, 13W/4L, $\text{Holm}_5 = .0895$); the same-run paired Lasso control above instead shows a directional full-joint-optimizer-bundle gain below the strict practical threshold. Neither result identifies a causal mechanism. The coverage extension includes all 17 regression datasets, but no cohort in this study is untouched confirmation material.

The -0.10 pp of the Poly2 row is a downstream-score gap under a fixed Ridge readout and is a different quantity from the -0.10 pp of Section 6, which is a span-R^2 difference against a width-matched quadratic on a different cohort. The two collide in rounding only and must not be quoted as one.

Finally, this protocol evaluates a Ridge readout only. We excluded Lasso because it emitted convergence warnings on a substantial minority of rows—21 of the 120 rows contributed by the four new datasets, and 81 rows overall—and a readout that does not converge cannot arbitrate between mappings. Because the downstream model here is Ridge alone, these results are unaffected by the boosting-library difference discussed in Appendix G. The *Lasso* column of Table 3 is a different quantity: the benchmark grid's Lasso readout over all 17 regression datasets, on which five of the six classical methods are positive but none exceeds 12.2.

The arm-neutral study, and why it is not carried here. The other retained attribution protocol fits the compression layer separately within each arm and scores all 37 datasets under two downstream regimes. Its tables are not reproduced. Its random-circuit contrast is the one place where the trained circuit does separate from the untrained one: $+3.508$ pp over all 37 datasets under a fixed linear readout ($p = .005$) and $+7.443$ pp on the 17 regression datasets ($p = .003$). Neither survives Holm correction over its forty contrasts, where the smallest adjusted value is 0.126 on this very comparison, and 86% of the $+7.44$ pp comes from SkillCraft alone; dropping that dataset leaves $+1.004$ pp on the remaining 16 at $p = .006$ with a 13W/3L record. The collapse is a property of the comparison rather than of the circuit, since on regression QFX stacks three circuits there while the random arm uses one. Fixed-linear regression therefore separates the circuit from a fixed random arm only under the separately fitted front end, not under the frozen one, and we do not generalize that result.

F Noise Sensitivity: Post-hoc Feature Perturbation and Transform-Time Circuit Noise

Post-hoc feature perturbation. The exact values in this paragraph are retained artifact measurements; the archive does not contain their producing driver, so they are auditable but not end-to-end regenerable. This study perturbs the *finished* feature vectors: the circuit is trained and the features extracted noiselessly, and a depolarizing-style channel of magnitude $p \in \{0, 0.001, 0.005, 0.01, 0.05\}$ is then applied to the extracted values. It therefore measures how tolerant the downstream models are to corruption of the readout representation, and is *not* a simulation of gate-local noise, readout error, or finite sampling; those channels act inside the circuit and are treated separately below. We evaluate 9 representative datasets (5 regression, 4 classification) with 5 seeds each (45 configurations per perturbation level). Two readings of "depolarizing-style" are available and we measured both. Shrinking each extracted value toward zero, $z \mapsto (1 - p)z$, is what a depolarizing channel does to a Pauli expectation. It moves the benchmark by only $+0.08$ pp in total because tree and linear estimators are invariant to monotone per-feature rescaling and the reported score is a maximum over them. We therefore apply the second reading, replacing each value independently with probability p by a draw from the maximally mixed outcome distribution, and report it here.

Under that channel the mean best-model score is 85.90 ± 0.47 at $p = 0$ and 82.69 ± 0.45 at $p = 0.05$, a relative reduction of 3.7%. Degradation tracks the task, not the sample size: the five regression datasets lose 5.58 pp on average and the four classification datasets 0.25 pp, a factor of twenty-two. Diabetes falls -20.0%, CPU Perf. -11.4% and Space GA -8.9%, while Sonar (-1.1%),

California (-1.0%), Yacht (-0.7%), Breast Cancer (-0.2%), Wine (0.0%) and Digits 5v8 ($+0.2\%$) barely move. Accuracy changes only when a corrupted feature pushes a prediction across a decision boundary, whereas R^2 responds continuously to any loss of fit, which accounts for the split without appealing to sample size. What separates the regression datasets from one another we do not claim to explain, since neither baseline R^2 nor sample size orders them: CPU Perf. starts at 94.08 and loses 10.69 pp, the largest absolute drop of the five, while California starts lower at 81.16 and loses 0.80 pp. For the majority of datasets the extracted representation therefore tolerates perturbation up to magnitude $p = 0.05$ with little loss. Because the perturbation is applied after extraction, this result speaks to the robustness of the features as inputs to the downstream model, and does not by itself establish a tolerable hardware error rate.

Transform-time circuit noise. The exact values in this paragraph are likewise artifact-only: the archive does not contain the transform-time noise driver and cannot regenerate them. The perturbation above acts on finished features. A separate sweep instead injects noise *inside* the circuit during feature extraction, so the compression layer and the variational parameters are trained and the features are read out under the noise channel rather than after it. We simulate with the PennyLane `default.mixed` density-matrix backend and vary four channels. These are gate depolarization at $p_{\text{gate}} \in \{0.001, 0.005, 0.01, 0.03, 0.05\}$, readout bit-flip at $p_{\text{ro}} \in \{0.01, 0.03\}$, and finite sampling at 1024 or 256 shots against the analytic default. Two combined settings pair $p_{\text{gate}}{=}0.005$ with $p_{\text{ro}}{=}0.01$ at analytic and at 1024 shots. The readout is fixed Ridge throughout, and each configuration is compared against Original, Poly2 and Nyström features computed on the same splits and seeds.

Density-matrix simulation costs 4^q rather than 2^q, and one dataset at $q{=}8$ consumed 57 CPU-hours for 11 configurations at 3 seeds. The sweep therefore covers five datasets rather than the full cohort, at one to three seeds each. Kin8nm was run at both $q{=}6$, the value this paper fixes, and $q{=}8$; those runs, Yacht, and CPU Perf. each carry the full 11 noisy configurations. Space GA carries 7, and Diabetes contributes only a noiseless reference. Only two of the five hold a noiseless margin over the classical references at all, so only those two give a crossover to locate.

Two readings follow, and they point in opposite directions. On the two datasets where QFX holds a noiseless advantage over every classical reference, that advantage survives gate depolarization at $p_{\text{gate}}{=}0.03$ and is lost between 0.03 and 0.05 on Kin8nm, at $p^*{=}0.042$ for $q{=}6$ and $p^*{=}0.039$ for $q{=}8$, where p^* is obtained by linear interpolation between the bracketing grid points. The qubit count therefore does not move the threshold materially: the Kin8nm margin over the best classical reference is $+9.39$ pp at $q{=}6$ and $+12.70$ pp at $q{=}8$. Yacht retains its margin across the entire sweep because that margin starts at 18.67 pp. On the remaining three the classical references already lead in the noiseless limit—CPU Perf. by 7.46 pp, Space GA by 5.18 and Diabetes by 3.62—so no crossover exists to measure and the sweep says nothing about them.

The channels are also unequal. Readout error is nearly free: at $q{=}6$ on Kin8nm, $p_{\text{ro}}{=}0.01$ gives 69.62 and $p_{\text{ro}}{=}0.03$ gives 69.50 against the noiseless 69.67. Finite sampling costs more but not enough to close the gap, at 67.53 for 1024 shots and 66.21 for 256. The combined settings land at 68.97 and 68.39, both still above the 60.28 of the best classical reference. Gate depolarization is the binding channel, and the p^* above is set by it alone.

We state the boundary of this result plainly. Five datasets at one to three seeds do not establish NISQ robustness, no result here was obtained on hardware, and only two of the five carry a noiseless margin for a crossover to be defined on. The figure $p^* \approx 0.04$ locates where a *uniform* depolarizing channel erases the margin in simulation; it is not an error budget for a device, because a real processor's error structure is gate-, qubit- and time-dependent in ways this channel does not model. We therefore report p^* as the answer to the question of where the simulated advantage ends and draw no deployment conclusion from it.

G Artifact scope and integrity

Practical and reproducibility boundary. The evidence supports no practitioner rule. Raw-feature XGBoost remains an external ceiling without its producing-source snapshot, non-row-local OpenFE covers only 4 of 17 regression datasets, and transfer reuses scored outcomes. The 2.0 pp analysis convention has no familywise operating characteristic, so failure to clear it does not exclude

all meaningful effects. The archive manifest records regeneration boundaries and uses null rather than inferred values for unknown provenance fields.

The anonymous `Supplementary.zip` contains the `qfx_v2` Python implementation and tests, protocol manifests, compact ledgers and reports, the ICLR experiment scripts and result summaries, a machine-readable dataset provenance manifest, environment metadata, and a hashed inventory. It excludes large dataset containers, model arrays, pickle files, caches, lock files, and raw or failed logs. It does not contain the transform-time noise driver of Appendix F or the q=6, 50,000-pair expressibility driver of Appendix N. It contains no expressibility driver at all, compatible or otherwise: the retained legacy diagnostic — an incompatible q=8, 5,000-pair script whose gradient-variance path fails silently under the pinned PennyLane version, and which could not reproduce the divergences reported here in any case — sits outside the bundle and is not packaged. Three intermediate files—`multi_seed_results_30.csv`, `lasso_combined_results.csv`, and `new_baselines_results.csv`—have no retained runnable producing drivers, so they can be audited as supplied artifacts but not regenerated end-to-end from the bundle. Trained circuit parameters are also absent. For all but one downstream estimator, settings are read from retained drivers. The exception is the ℓ_1-penalized logistic regression behind the classification *Lasso* column of Table 3. Its retained rows are the union of `lasso_combined_results.csv` and the classification Lasso rows in `multi_seed_results_30.csv`. Candidate-estimator fitting on the retained splits and seeds identifies `LogisticRegression(penalty='l1', C=1.0, solver='liblinear')`, which matches every rerunnable classification score within 5×10^{-7} and is listed in Table 7. This setting is inferred from outputs rather than preserved in a producing driver.

The classification ℓ_1 record is incomplete, which affects one column of Table 3. The combined file contributes 12 current datasets; the older multi-seed file adds partial rows on HAR, ISOLET, Jasmine, Semeion, and Volkert. Madelon, Mfeat-Factors, and Mfeat-Pixel have no ℓ_1 fit. After averaging available seeds within dataset and then datasets, coverage is 17 for Original, 16 for PCA, 15 for Poly2, RFF, and Nyström, 13 for OpenFE, and 12 for QFX and every other quantum arm. These reductions reproduce all printed classification-Lasso scores. The score cells are not like-for-like and receive no emphasis. For the rank column, each dataset uses descending average-tie ranks over the displayed methods available on that dataset, followed by a mean over the datasets available for each method. This reproduces every printed rank; because the available method set and dataset cohort vary, the mean across the 12 method ranks is 6.15 rather than the 6.50 of a complete 12-method ranking, and the cells likewise receive no emphasis.

The reconstructed cohort and the boosting library. Seven of the 37 datasets lack raw per-seed files and were reconstructed with the same pipeline and seeds. The available reconstruction first used xgboost 1.7.1; the reported boosted cells use the pinned 2.1.4 build. Holding feature matrices fixed, Ridge, Lasso, Linear, Logistic, and Tree cells are invariant across these builds, whereas XGB cells move by -0.62 to $+0.53$ pp. Of 39 stored anchor cells on these datasets, 31 reconstructed cells lie within one anchor standard deviation and eight do not: four on Residential (Original, PCA, Poly2, and OpenFE, higher by 2.4–3.2 pp), two on Madelon (OpenFE lower by 3.8 pp and QFX by 3.4), and two on UJIIndoorLoc (Base by 0.20 pp and QFX by 0.13). Anchored cells outside the reconstruction differ by at most 0.23 pp. A few logistic and ordinary-least-squares cells also retain one-to-two-sample discrepancies regardless of xgboost build; Residential has the largest residual because its least-squares fit is underdetermined. The reported field uses the reconstructed, pinned-build values and discloses these anchor disagreements without reconciling them.

H Dataset Details

Table 5 lists the 37 benchmark datasets (17 regression and 20 classification) with task, roster-row count before lane-specific caps, feature count, and class count. Every named benchmark cohort is a subset of these rows. Regression uses R^2 and classification uses accuracy in the augmentation grid; lanes with different metrics are identified in Section 3. The archive's machine-readable provenance manifest records recovery status and checksums where evidenced; unavailable source, version, or target fields are null rather than inferred. The augmentation grid caps the full dataset at 10,000 rows before its 70/30 split, with classification stratified where feasible; where the cap binds, this yields at

Table 5: The 37-dataset roster: 17 regression datasets, scored by R^2 in the augmentation grid, and 20 classification datasets, scored there by accuracy. Row counts precede lane-specific caps and splits; E4 uses the full listed roster rows, whereas the augmentation grid applies its 10,000-row cap before splitting. Other evidence lanes use the metrics declared in Section 3. Names and dimensions are verified from the available roster; exact source/version/target/checksum provenance is unavailable.

Dataset	Task	Roster rows (pre-cap)	#Features	#Classes
Regression				
Mercedes-Benz	REG	4,209	356	—
Abalone	REG	4,177	8	—
Ailerons	REG	13,750	40	—
Bank32nh	REG	8,192	32	—
California	REG	20,640	8	—
CPU Perf.	REG	209	7	—
CT Slice	REG	53,500	384	—
Diabetes	REG	442	10	—
Insurance	REG	1,338	6	—
Kin8nm	REG	8,192	8	—
Protein	REG	45,730	9	—
Residential	REG	372	107	—
SkillCraft	REG	3,338	18	—
Space GA	REG	3,107	6	—
Tecator	REG	240	124	—
UJIIndoorLoc	REG	10,000	454	—
Yacht	REG	308	6	—
Classification				
Adult	CLS	45,222	14	2
Breast Cancer	CLS	569	30	2
Churn	CLS	7,043	20	2
Digits 5v8	CLS	356	64	2
Glass	CLS	214	9	6
HAR	CLS	10,299	561	6
ISOLET	CLS	7,797	617	26
Jasmine	CLS	2,984	8	2
Madelon	CLS	2,600	500	2
Mfeat-Factors	CLS	2,000	216	10
Mfeat-Pixel	CLS	2,000	240	10
Parkinsons	CLS	195	22	2
Segment	CLS	2,310	19	7
Semeion	CLS	1,593	256	10
Sonar	CLS	208	60	2
Vehicle	CLS	846	18	4
Volkert	CLS	58,310	147	10
Vowel	CLS	990	12	11
Wine	CLS	178	13	3
Wine Quality	CLS	6,497	11	7

most 7,000 training and 3,000 test rows. Other evidence lanes follow their own sampling. E4 uses the full roster for its sixteen datasets before its 70/30 split.

I Statistical Significance Details

The following 37-dataset tests pool regression and classification and are exploratory diagnostics; they support neither a task-specific inferential verdict nor a common-scale effect claim. Wilcoxon signed-rank tests with Holm–Bonferroni correction for 8 pre-specified pairwise comparisons (QML-Direct, FC-VQC, QFX-QO, PQK, CoVar, ZZFeatureMap, Poly2, OpenFE) give adjusted $p < 10^{-9}$ for QML-Direct and FC-VQC, $p < 10^{-8}$ for QFX-QO, $p < 10^{-3}$ for PQK and CoVar, $p < 10^{-2}$ for ZZFeatureMap, $p = 0.96$ for Poly2, and $p = 0.42$ for OpenFE. The correction applies only to these eight comparisons and does not establish equivalence for Poly2 or OpenFE. The three ablation

comparisons—QFX vs. Base, Wide-Only vs. Deep-Only, and All-Stack vs. QFX—were selected after inspection and are exploratory.

FT-Transformer sub-cohorts. The FT-Transformer column of Table 2 aggregates two implementations of one architecture family: one covers 30 datasets and the other covers 7. A mean within that column cannot isolate an implementation effect. The rank column instead computes each dataset's rank among the 25 methods before averaging ranks. The two sub-cohorts place QFX similarly relative to FT-Transformer: $+9.06$ pp on the 30-dataset cohort (25W/5L) and $+8.49$ pp on the 7-dataset cohort (6W/1L). This is not an isolated measurement of implementation. Implementation, split protocol, and target standardization vary together across the two cohorts, so the agreement does not establish equivalence. Across all 37 datasets, FT-Transformer exceeds QFX on six; the widest margin is Abalone at -22.39 pp.

TabNet's regression mean. TabNet's regression entry in Table 2 reads -24.07 ± 215.36, and the score and rank columns of that row point in opposite directions: on score it sits below QML-Direct, on rank well above it. Both are correct, and the reason is that two of its seventeen regression fits diverge—UJIIndoorLoc at $R^2 \times 100 = -967.9$ and SkillCraft at -396.2—while the remaining fifteen are positive. An unbounded-below score averages those two into the mean; a rank cannot, because each per-dataset rank is ordinal. The median across the seventeen is 52.5, and clipping R^2 at zero gives 56.2. We print the unclipped mean in Table 2 so that the failures are visible rather than absorbed, and note here what it is measuring. TabNet's twenty classification datasets contain no diverged fit and need no such qualification.

Comparators evaluated only in part. Two current comparators are available only on partial cohorts; positive gaps favor QFX. CatBoost (Prokhorenkova et al., 2018) gives $+1.29$ pp with a 14W/9L record over 23 datasets and is not significant. LightGBM (Ke et al., 2017) gives $+3.25$ pp on a 16W/6L/1T record over the same 23, with uncorrected p=0.039; the Holm adjustment was computed over the three partial-cohort comparators this lane evaluated and gives p_{adj}=0.118. Both arms are incomplete comparisons and do not enter the 37-dataset ranks; the complete gradient-boosting record on the sealed margins grid is Table 6 below.

The same k=25 field also yields a post-hoc diagnostic against QFX, carried by the two columns beside the OVR ranks of Table 2: a pooled Nemenyi test (Demšar, 2006) over 37 datasets and 25 methods puts the critical difference at 6.26 average-rank units at $\alpha = 0.05$, and QFX exceeds it against eight of the twenty-four other arms. It separates from no classical baseline, and from none of PVQNN, QRC, QKS or PQK; the eight it does separate from all sit at OVR 14.46 or worse. Those two columns compare each arm against QFX alone and license no statement about the remaining pairs. Because this diagnostic pools task types, it is exploratory rather than task-specific inference.

Complete untuned gradient-boosting record on the sealed margins grid. Separately from the partial-cohort comparators above, run `practical_ceiling_20260818` scores XGBoost, LightGBM, and CatBoost on the raw features of all 17 regression datasets under the sealed fixed-margins protocol of Section 5: the 10,000-row cap and 70/30 split, five seeds, one test-set score per cell, and split hashes equal to the fixed-head grid on all 85 dataset–seed cells. XGBoost uses the fixed settings of Table 7; LightGBM 4.6.0 and CatBoost 1.2.10 use library defaults with only the seed set, so all three are untuned. Table 6 lists seed-averaged $R^2 \times 100$ beside the fixed Lasso-T1 Original and QFX scores carried over unchanged from that lane. The record is an external practical ceiling: the ensembles differ from every feature-map arm in both features and estimator, belong to no contrast family, and carry no decision rule. Against fixed Lasso-T1 Original the median gaps are $+10.65$, $+9.62$, and $+9.90$ pp with 14–3–0, 13–4–0, and 13–4–0 records for XGBoost, LightGBM, and CatBoost; against fixed Lasso-T1 QFX they are $+0.76$, $+1.31$, and $+1.12$ pp with 11–6–0, 9–8–0, and 10–7–0. Medians and win–loss–tie records are the primary descriptors because R^2 is unbounded below and two reference cells diverge (California -48.06, SkillCraft -566.51); means carry that caveat. The ceiling is not uniform: all three ensembles trail Original on Diabetes, Residential, and Tecator, with a fourth loss on CPU Perf. for LightGBM and on Mercedes-Benz for CatBoost, and each trails QFX on six to eight datasets. The ledger SHA-256 is `3b32f23341ec33081fb43e7986360249 8894f3758862538878225e5ec96cc89d`; the result-record SHA-256 is `7cd8aea9fe9a82c7b7b8bcb9385661d7e1d8b665abf8dfdd750f66e406e37b97`.

Tuned gradient-boosting record on the same grid. Run `tuned_tabular_baselines_20260922` repeats the three ensembles with hyperparameters selected per dataset–seed–model cell by a 40-trial tree-structured Parzen estimator search over three inner folds of the outer-training rows only (scaler refit per fold; XGBoost depth 3–10, LightGBM 8–256 leaves, CatBoost depth 3–8, with the learning-rate, regularization, and subsampling ranges fixed in the lock), refit once on the full outer-training rows, and scored once on the sealed test split; split hashes equal the fixed-head grid and the untuned ledger on all 85 cells. The lock was finalized before scoring; its one amendment, after a machine shutdown, changed only the worker partition, with the hash-stamped driver and analyzer unchanged. Tuning raises the seed-averaged medians to 82.69, 84.26, and 85.28; the median per-dataset tuning gain is $+1.53$, $+1.84$, and $+1.66$ pp (17–0–0, 17–0–0, and 14–3–0). Against fixed Lasso-T1 Original the median gaps are $+11.63$, $+11.46$, and $+11.92$ pp with 14–3–0, 13–4–0, and 16–1–0 records for XGBoost, LightGBM, and CatBoost; against fixed Lasso-T1 QFX they are $+2.11$, $+1.68$, and $+2.03$ pp with 15–2–0, 14–3–0, and 14–3–0. The tuned ceiling is external and descriptive in the same sense as the untuned one: it belongs to no contrast family, carries no decision rule, and changes no verdict. It trails QFX on Residential and Tecator for all three ensembles and on CPU Perf. for LightGBM and CatBoost, and trails Original on Tecator for all three, on Diabetes and Residential for XGBoost and LightGBM, and on CPU Perf. for LightGBM. Among the eight columns of Table 6, QFX has median rank 7, first on Residential and Tecator and second on CPU Perf. Tuned tabular learners beyond gradient boosting were not run. The ledger SHA-256 is `c999c7620a58c3435d8d4e22e8bfeb7b2476d8a0f3ad711c5ca034155790bb10`; the result-record SHA-256 is `311a06fdc99c1280d8aede3c787cf8a160056a0ec31e6605cfcc32a7256e71c5`.

Table 6: Untuned and tuned gradient-boosting ceilings on the sealed margins grid, runs `practical_ceiling_20260818` (untuned columns) and `tuned_tabular_baselines_20260922` (tuned columns: hyperparameters chosen per cell by a 40-trial search on three inner folds of the outer-training rows, one score call per cell). Entries are $R^2 \times 100$ averaged over five seeds on raw features; the Original and QFX columns are fixed Lasso-T1 scores carried over unchanged from the margins lane for context. The record is descriptive and outside every contrast family. The three ensembles differ from the two feature-map columns in both features and estimator, so cells are not comparable across that boundary and no entry carries bold or underline emphasis. The two diverged cells, California -48.06 and SkillCraft -566.51, both sit in the Original column, so the mean row is unusable only for Original-referenced comparisons; the QFX column carries no diverged cell and its mean is read directly. Against that column the paired mean gaps are $+7.50$ pp for XGBoost, $+5.82$ for LightGBM and $+9.17$ for CatBoost, and the paired median gaps are $+0.76$, $+1.31$ and $+1.12$; for the tuned columns the paired mean gaps are $+10.40$, $+9.23$ and $+11.22$ pp and the paired median gaps $+2.11$, $+1.68$ and $+2.03$.

	Untuned			Tuned				
Dataset	XGBoost	LightGBM	CatBoost	XGBoost	LightGBM	CatBoost	Original L/T1	QFX L/T1
Mercedes-Benz	55.40	53.32	53.18	56.51	55.51	56.20	53.23	54.39
Abalone	33.22	30.65	31.95	35.09	35.04	35.35	28.18	34.25
Ailerons	82.36	84.61	85.48	82.49	84.86	85.28	−0.02	−0.02
Bank32nh	54.69	55.35	57.13	56.21	56.52	57.13	48.30	48.94
California	81.87	83.12	84.10	84.04	84.49	84.07	−48.06	70.06
CPU Perf.	91.37	59.36	85.75	92.61	69.19	91.57	83.59	92.23
CT Slice	96.14	97.59	97.59	98.66	98.59	98.21	85.47	88.00
Diabetes	34.74	37.14	41.19	45.19	45.97	47.83	46.78	35.31
Insurance	85.03	84.00	84.28	86.01	85.84	86.29	74.38	84.27
Kin8nm	71.06	77.31	85.93	82.69	84.26	89.17	40.59	67.02
Protein	48.69	51.36	54.66	56.98	57.49	57.46	28.22	35.05
Residential	96.28	93.94	96.77	96.76	96.82	97.44	97.33	97.94
SkillCraft	60.76	60.42	61.09	62.39	62.06	62.65	−566.51	62.01
Space GA	67.02	68.88	70.05	70.50	69.79	71.71	40.85	61.49
Tecator	98.69	97.16	96.77	99.26	98.96	99.00	99.56	99.64
UJIIndoorLoc	98.19	99.00	98.80	99.17	99.37	99.08	93.60	97.68
Yacht	99.57	93.24	98.74	99.80	99.78	99.83	62.71	99.31
Median	81.87	77.31	84.28	82.69	84.26	85.28	48.30	67.02
Mean	73.83	72.15	75.50	76.73	75.56	77.55	15.78	66.33

J Downstream Model Analysis

One selection-free measurement of the benchmark grid is recorded here. Scoring seven feature transformers per task against every downstream estimator in turn, taking no maximum anywhere, is the only place the unpenalized least-squares readout is scored on its own. Under it QFX collapses to a regression $R^2\times100$ of -1047, against 68.31 for single-circuit Base, because the q=6, 69-feature stack leaves that fit ill-conditioned. The benchmark margin is therefore specific to *penalized* linear models, and that collapse is what makes the "fixed-penalty" qualifier the body attaches to its linear readouts a measured restriction rather than a convenience.

Downstream performance aggregates two distinct effects: tree-based performance, where classical nonlinear methods are already strong, and linear-model performance, where quantum features are most distinctive. To separate them we widen the field to all 20 feature transformers—the five end-to-end predictors take no downstream model and are excluded throughout—and restrict the downstream model to one estimator family before ranking, over the same 37 datasets, 5 seeds and splits. The tree family is XGBoost or a depth-5 decision tree; the linear family is Ridge, Lasso or Linear Regression on regression together with the two ℓ_2-penalized logistic readouts at C=1.0 and C=0.1 on classification (Table 7), with the ℓ_1 logistic fit excluded because its retained cohort is incomplete (Appendix G). Each family takes the maximum over its own estimators for every transformer–dataset–seed triple before averaging the 5 seeds, so both are best-of-k views under the benchmark protocol and are not commensurable with the XGBoost-only regime below, which fixes one estimator in advance. Ranks are per-dataset ranks among the 20 transformers with ties averaged, meaned over the 37 datasets; scores pool the 17 regression and 20 classification datasets into one mean and are read down a regime, not across the two task types.

Tree-based downstream models. With tree-based readouts, QFX ranks sixth among the 20 transformers at 10.19, behind four of six classical transformers and Base. Poly2 leads at 6.43; mean scores are 80.91 for Poly2 and 80.33 for QFX. These comparisons do not identify the mechanism.

Linear-restricted downstream models. When the downstream model is constrained to linear estimators, QFX attains a mean linear-restricted rank of **6.55** across the 37 datasets, placing third among the 20 feature transformers and ahead of every transformer outside its own variant family, the best of those being Nyström at 7.18; Poly2 drops to rank 11.11, OpenFE to 13.65, and Original to 14.55, reversing the tree-restricted ordering. The mean linear-restricted score of QFX is 78.99 on all 37 datasets, leading the non-quantum transformers, where Nyström reaches 76.96, OpenFE 63.74 and Original 53.19. The mean linear-restricted score of Poly2 on the full benchmark drops to -9.64×10^5 because collinearity-induced negative R^2 on a subset of regression datasets dominates the average, whereas QFX does not collapse on those datasets. The contrast between the 6.55 linear-restricted rank and the 10.19 tree-restricted rank does not by itself attribute the difference to a learned quantum basis: Section 5 shows that fixed-penalty readouts are sensitive to transformed-block scale, and Section 6 places the deployed circuit's readout variance inside the degree-2 monomial span of its compressed input. Two variants rank lower still, All-Stack at 5.70 and Deeper at 6.26.

Best downstream model selection. The best-model protocol selects the estimator with the highest test score and has no separate validation partition. We count per transformer what that protocol picks, so that the circuit's frequency is read against the field rather than on its own: QFX selects a non-tree readout on 22 of 37 datasets and on 12 of 17 regression datasets, against 6 to 19 of 37 for the six classical transformers. This optimistically selected frequency is descriptive and consistent with the fixed Ridge/Lasso results; it does not itself confirm a mechanism.

XGBoost-only ranking. A third regime ranks all feature transformers with XGBoost fixed in advance as the sole downstream model, so its numbers are free of downstream-head selection within this fixed-estimator view. Poly2 leads at 6.32, OpenFE at 6.64, and PCA at 7.03, while QFX ranks 10.42. The quantum arms and weaker classical ones interleave: QFX is ahead of five of six fixed-circuit maps and behind QRC at 10.32, while RFF at 11.80 trails all six. Together with the fixed Ridge/Lasso table (Table 3), this shows downstream-model dependence; it does not identify the mechanism. Set the three regimes beside one another and none of them can be quoted as the standing of the feature map: on the same 20 transformers, datasets, seeds and splits, QFX moves from 10.42

with XGBoost forced, to 10.19 in the tree family, to 6.55 in the linear family—seventh of twenty, then sixth, then third.

Selection-free and best-of-k views. Table 3, the unpenalized-readout measurement above and the XGBoost-only regime each fix a single estimator in advance from Table 7, so their scores are free of downstream-head selection within the corresponding fixed-estimator view. This limited property does not remove benchmark adaptation in the feature-map design described in Appendix L. Table 2 and the tree- and linear-family regimes instead maximize over estimators scored on the test split and are optimistic as absolute estimates, so the two kinds of number must not be read at equal weight. All 20 feature transformers share the estimator set, splits and seeds, but the premium is not uniform: on regression it spans 0.76–3.40 pp, and QFX collects 0.45 pp more of it than Poly2 against the 0.02 pp that separates them in Table 2, so close margins in the best-of-k view carry no ordering. The five end-to-end predictors receive no premium at all, which averages $+1.57$ pp on regression and $+0.97$ pp on classification, so Table 2 favors the feature arms over those five rather than the reverse. Where the two views disagree we treat the fixed-estimator view as primary.

K Augmentation and Downstream Hyperparameters

Table 7 lists the retained augmentation settings and the downstream estimator defaults shared by feature-transformer arms. It does not describe the controlled CKA protocol, which is specified in Appendix B. Tuned MLP is the per-dataset exception described below.

The 5 nominal seeds index the train/test split and circuit initialization. The single-circuit strategy-selection path also has the additional run-to-run variation quantified in Appendix L.

Downstream estimator sets. Regression uses five estimators (XGBoost (Chen & Guestrin, 2016), Ridge, Lasso, Linear Regression, Decision Tree) and classification uses four (XGBoost, ℓ_2-penalized logistic regression at C=1.0, the same estimator at C=0.1, Decision Tree). The downstream estimator settings listed above were fixed before any benchmark run and applied identically to every feature transformer and every dataset, with no per-dataset or per-method tuning. Lasso's α=0.01 is not the scikit-learn default (α=1.0); it was chosen once, in advance, and held constant throughout. That does not extend to the rest of Table 7: of the twelve design settings a practitioner has to choose, eight were fixed after inspecting benchmark results, and of the remaining four only two — this estimator set and the seed set — are cleanly pre-fixed, since the mutual-information pre-selection width has no surviving pre-benchmark record and the training-sample formula was chosen on Yacht and Adult, both benchmark datasets. Appendix L separates all twelve by provenance one setting at a time.

Reading the linear column names on classification. No ordinary least-squares estimator—Ridge, Lasso, or plain Linear Regression—is fitted on a classification task. Where a classification panel carries the headings *Ridge*, *Logistic*, or *Lasso* (Table 3), those names abbreviate penalized logistic regressions. *Ridge* and *Logistic* are the same ℓ_2 estimator at C=1.0 and C=0.1, so their gap measures regularization-strength sensitivity, not a model-family difference. *Lasso* is ℓ_1-penalized logistic regression at C=1.0. It contributes the classification *Lasso* column of Table 3 but is excluded from the four-estimator maximum in Table 2. Its configuration is absent from the retained drivers (Appendix G); the table records a reconstruction that matches available per-seed outputs within 5×10^{-7}. It is not a recovered record.

Quantum baseline specifications. All six fixed-circuit baselines are sized by $q = \min(n, 8)$. That is a normalization we impose so the six sit on one grid, not a rule any of them publishes; their original works size their registers differently and in ways that do not transfer to a common tabular benchmark, and where a paper reports a fixed register we do not adopt it. The same caveat applies more sharply to the FC-VQC arm (Su et al., 2026), which we implement as one fully-connected layer into a single register of at most 12 qubits followed by one 1W+2D circuit. That is a monolithic configuration, and the architecture it is named for is not: a reader should treat our FC-VQC column as this configuration under our sizing rather than as a reproduction of that work, and the retained $35/37$ record as scoped accordingly. They share the benchmark preprocessing: Quantile-Transformer scaled to $[0.05, \pi - 0.05]$. A multi-basis circuit therefore emits $3q + (q-1)$ features—31

Table 7: Augmentation settings for QFX and downstream defaults shared by feature-transformer arms. The controlled CKA protocol is specified in Appendix B; Tuned MLP instead uses a disclosed 20-trial per-dataset search. Classification aliases denote penalized logistic regressions, and the ℓ_1 setting is reconstructed from retained outputs (Appendix G).

Component	Hyperparameter	Value
Augmentation circuit	Number of qubits (q)	6
	Architecture	1 Wide + 2 Deep (1W+2D)
Augmentation training	Optimizer / loss	Adam / hybrid loss, Equation (4)
	Learning rate	0.01
	Batch size	64
	Maximum steps	100
	Early stopping	patience 20; retain halting-step parameters
	Learning-rate decay	factor 0.95 every 10 steps
	Gradient clipping	configured at 1.0; absent from retained fit loop
	Original features	concatenate $[x, f(x)]$
	Compression activation	$\tanh$ if $\rho_{\text{enc}} \geq 10$; cosine otherwise
Multi-init seeds	Seed set	{42, 123, 456, 789, 2024}
	Init strategies (regression)	Identity, Random, PCA
	Init strategy (classification)	Identity only
Augmentation loss	Correlation penalty λ	0.3
	Sigmoid gate	$\alpha = \text{sigmoid}(0.5 \cdot (\rho_{\text{enc}} - 8))$
Preprocessing	QuantileTransformer $n_{\text{quantiles}}$	$\min(1000, N_{\text{train}})$
	Feature scaling range	$[0.05, \pi - 0.05]$
Data splitting	Full-dataset cap, then split	10,000 rows, then 70/30
	Maximum outer train/test	7,000 / 3,000 rows when capped
	Circuit-fitting subsample	$\min(\max(700, 5\sqrt{N}), 5000)$ training rows
Simulation	PennyLane version	0.38.0
	Backend	`default.qubit`
	Differentiation method	Backpropagation
Downstream (regression)	XGBoost	`n_estimators`=100, `max_depth`=5, `learning_rate`=0.1
	Ridge	$\alpha = 1.0$
	Lasso	$\alpha = 0.01$, `max_iter`=5000
	Linear Regression	none
	Decision Tree	`max_depth`=5
Downstream (classification)	XGBoost	`n_estimators`=100, `max_depth`=5, `learning_rate`=0.1
	ℓ_2 logistic (C=1.0)	`lbfgs`, 1000 iterations; "Ridge" alias
	ℓ_2 logistic (C=0.1)	`lbfgs`, 1000 iterations; "Logistic" alias
	ℓ_1 logistic (C=1.0)	`liblinear`; reconstructed "Lasso" alias
	Decision Tree	`max_depth`=5

at the modal q=8, 27 on CPU Perf. (n=7) and 23 on the three n=6 datasets. **PVQNN**: an ensemble of R_{PVQNN}=5 diverse, *fixed* quantum circuits with different architectures—RY/RX/RZ encoding and linear/ring/star CNOT topologies—under that readout, for $5\,[3q+(q-1)] = 155$ quantum features at q=8. **QRC** (Fujii & Nakajima, 2017): a quantum-reservoir-computing-inspired single circuit with RY encoding followed by 8 layers of fixed random unitaries in an alternating brick-wall CNOT pattern, under the same $3q+(q-1)$ readout. **QKS** (Wilson et al., 2018): a QKS-inspired bank of R=10 random parameterized circuits with data-dependent RY encoding plus fixed random shifts, ring CNOT topology, and Z-basis-only measurements, producing $10q = 80$ features at q=8; the 100 training steps listed in Table 7 refer to the classical readout, not quantum parameters. None of these six quantum feature map baselines involve gradient-based quantum parameter optimization (Mitarai et al., 2018)—only the classical readout model is trained. PQK, the projected quantum kernel (Huang et al., 2021), and ZZFeatureMap (Havlíček et al., 2019) are named-reference adaptations under the common $q = \min(n, 8)$ grid. CoVar is the benchmark's custom covariance-inspired circuit recovered from retained code, not a reproduction of a cited method.

Deep tabular baseline specifications. **Tuned MLP**: separately for each dataset, the scikit-learn (Pedregosa et al., 2011) MLP's architecture and regularization are selected by a 20-trial Optuna search on an inner $80/20$ split of the seed-42 training data. The search space is 1 or 2 hidden layers, width $\{64, 128, 256, 512\}$, an ℓ_2 penalty log-uniform on $[10^{-6}, 10^{-2}]$, initial learning rate log-uniform on $[10^{-4}, 10^{-2}]$, and batch size $\{64, 128, 256\}$. That dataset-specific winner is then refit unchanged on all five reported seeds. **TabNet** (Arik & Pfister, 2021): the `pytorch_tabnet` implementation uses the published architecture defaults with a 300-epoch cap and early-stopping patience 40; the retained evidence does not record the package version. **FT-Transformer** (Gorishniy et al., 2021): as Appendix I discloses, one implementation covers 30 datasets and a second covers 7.

All three share the benchmark's common preprocessing path: inputs are standardized with `StandardScaler`, regression targets are standardized, and each seed caps the full dataset at 10,000 rows before the 70/30 split, with classification stratified where feasible. Where the cap binds, every arm receives at most 7,000 outer-training rows. The remaining differences are explicit: these three select no downstream model, the two FT-Transformer cohorts differ in target standardization and split parameters, and that row spans two implementations (Appendix I). The deep baselines are reference implementations under this fixed budget rather than tuned upper bounds.

L Provenance of the design settings

The design settings of QFX were not all fixed at the same point, and some were selected from the same 37 datasets used for evaluation. Discovering and evaluating a rule on one benchmark inflates its apparent quality, so this section separates the settings by provenance.

Provenance. Twelve practitioner-facing choices fall into three provenance classes. The downstream estimator set and its settings, including Lasso's $\alpha{=}0.01$, and the seed set $\{42, 123, 456, 789, 2024\}$ are recorded as fixed before the benchmark. The mutual-information preselection width $\tilde{n}{=}\min(n, 64)$ is architectural rather than tuned, but no pre-benchmark record survives, so its chronology is not established. The training-sample formula $\min(\max(700, 5\sqrt{N}), 5000)$ was chosen on Yacht and Adult and not swept on the 37-dataset grid.

The remaining eight choices are *in-sample*: they were selected after inspecting results from datasets used for evaluation. The qubit count $q{=}6$ is the benchmark-adaptive primary setting; the retained record does not establish that either noncommensurate sweep in Appendix M uniquely selected it. The diversity weight $\lambda{=}0.3$ and activation threshold $\rho_{\mathrm{thr}}{=}10$ come from 10-dataset, 5-seed sweeps, with the former reported in Appendix B. The unswept loss-gate center $\rho_{\mathrm{enc}}{=}8$ was defined as $\rho_{\mathrm{thr}}{-}2$ and inherits that provenance. The 1W+2D architecture and task-aware circuit count—three circuits for regression and one for classification—use the 37-dataset ablation in Table 2. The 100-step budget uses a 21-dataset, 5-seed sweep that resolved no budget effect, and the compression configuration uses a 6-dataset, 5-seed grid. These eight choices are adaptive, not externally confirmed.

The evidence grids just named are read against the transformer's own reproducibility floor, which we measure next.

Run-to-run variation. Fitting the transformer twice with identical arguments does not always give identical features. The single-circuit initialization contest is not fully pinned by the seed argument because a reseeded global state affects its subsample. Across 20 same-configuration comparisons the gap is at most 1.32 pp and exceeds 0.5 pp on three; the spread falls on SkillCraft, while Breast Cancer is identical. Averaging over seeds shrinks this variation, but a per-cell spread is a dispersion measure, not a standard error. We therefore do not attribute differences below roughly 1.3 pp to a design setting on close contests.

M Qubit count under two readout protocols

To assess sensitivity to qubit count, we conduct an ablation study varying $q \in \{2, 4, 6, 8, 10\}$ across 7 representative datasets (4 regression, 3 classification) with 5 seeds each (175 runs total). That sweep reports mean best-model scores and is measured under the protocol of Table 7.

Key findings. Under the seven-dataset, five-seed best-model protocol, the mean spread across qubit counts is 0.82 pp, with a maximum of 2.53 pp on Diabetes; the mean spreads are 0.64 pp for classification and 0.95 pp for regression. These figures describe only this estimator-selected cohort. Here q=6 is best on none of the seven datasets: q=2 takes four, and q=4, q=8, and q=10 one each, crediting a tie to the smallest q that attains it. In the separate fixed-Ridge sweep below, which changes cohort and seed count, q=2 loses to q=6 on 11 of 11 datasets by 23.23 pp. The two sweeps therefore do not establish general robustness to qubit count or isolate a circuit-capacity effect; the adopted q is not selected from the best-model sweep.

Protocol-dependent qubit sensitivity. The sweep above maximizes over estimators on seven datasets. A second sweep fixes Ridge over 11 nonlinear regression datasets at 3 seeds, 165 cells in all; the eleven are the nine nonlinear regression datasets of Table 4 plus Bike Sharing and Year Prediction, which lie outside the main benchmark. It changes the estimator protocol, cohort, task composition, and seed count simultaneously, so it is a complementary measurement rather than a direct attribution test. We report it descriptively because it is the only grid that records accuracy, simulation cost, and the head-to-head result against q=6 together. Accuracy figures from it are means of $R^2 \times 100$ over all cells and cost figures are mean wall-clock seconds over those same cells, while average rank is taken over the five qubit counts with the dataset as the unit, so it runs from 1 to 5.

The cohorts show different sensitivity. Under best-model selection, $q = 2$ is highest on four of seven datasets and the mean spread is 0.82 pp. In the 11-dataset fixed-Ridge cohort, $q = 2$ loses to $q = 6$ on all 11 by 23.23 pp on average, with negative cells on SkillCraft and California. Because the cohort and seed count also change, this is not a decomposition of the seven-dataset flatness and does not identify a representation-capacity mechanism.

At the top of the range accuracy favors the larger circuit, but by less than the measurement resolves. q=8 takes the best average rank at 1.64 and beats q=6 on 10 of 11 datasets, yet by 0.90 pp on the mean, inside the run-to-run floor of up to 1.32 pp measured in Appendix L, and at 2.7× the simulation cost of q=6, which averages 86.6 s a cell; q=10 adds nothing over q=8 and costs 11× what q=6 does. These two noncommensurate sweeps provide descriptive sensitivity evidence, not an isolated cost–performance selection or validation of q=6; q=6 remains the benchmark-adaptive primary setting.

N Expressibility–Trainability Discussion

We observe empirically that 1W+2D is not distinguishable from the 0W+3D, augmentation data-re-uploading or larger 2W+3D variants under the 100-step budget.[1] Their overall scores in Table 2 differ by less than the run-to-run floor measured in Appendix L, so we read none of those gaps as separating one architecture from another. We characterize all five augmentation architectures with expressibility and trainability diagnostics below.

Expressibility measurement. The exact values below are retained artifact measurements. The archive has no compatible q=6, 50,000-pair producing driver, so these values are auditable from the supplied artifact but not end-to-end regenerable. Following Sim et al. (2019), we quantify expressibility via the Kullback–Leibler (KL) divergence between the circuit's fidelity distribution and the Haar-random fidelity distribution, using 50,000 random parameter pairs per cell. Lower KL divergence indicates higher expressibility, reflecting closer coverage of the full unitary space. Two parameterizations have to be kept apart: the available diagnostic build parameterizes the all-to-all CR_X layer per pair, whereas the evaluated model shares one angle per control qubit, so their counts grow quadratically and linearly in q respectively, and every figure quoted here uses the evaluated parameterization. Measured that way over 15 independent cells, these diagnostics separate the variants where the downstream scores do not, and they separate exactly one. 3W+0D (Wide-Only) sits at a KL divergence of 51.826 ± 3.678 ($\times 10^{-4}$), an order of magnitude above 1W+2D

[1] Circuit training has patience 20 and carries forward the parameters at halting rather than the best-loss parameters. Against matching 100-step cells, a 400-step augmentation +Reupload probe leaves 57/100 unchanged; 24 worsen and 19 improve. Changes range from -7.692 to $+9.259$ pp, so the probe does not support a monotone budget claim.

at 4.849 ± 1.70, 0W+3D at 4.955 ± 1.04, augmentation +Reupload at 3.978 ± 0.986 and 2W+3D (Deeper) at 1.322 ± 0.305. Scoring exactly Haar-random circuits under the identical protocol over 500 trials puts the estimator's finite-sample floor at a median of 1.285 and a p95 of 2.223, so 2W+3D lies at that floor and saturates the measure rather than leading it. Of the three architectures that sit above the floor and below 3W+0D, every pair has overlapping $\pm$SD intervals, so the diagnostic orders none of them against the others. The single architecture it pulls away from the field is excluded from the evaluated downstream set, which is the sense in which it records what the adopted architecture was *not* chosen on.

3W+0D is the KL outlier; its gradient variance, 9.855 ($\times 10^{-2}$), is not distinctive, sitting beside the deployed 1W+2D at 9.568 and well below augmentation +Reupload at 22.122 in the same units. 2W+3D saturates the KL estimator's Haar floor and has the lowest gradient variance of the five.[2] The deployed parameter counts differ, but the retained sources do not establish a convention-matched numerical count for this comparison. Because the downstream scores of Table 2 do not separate the two, the reported trade-off is descriptive rather than a matched-budget test or a diagnosed optimization mechanism.

Gradient variance analysis. Following McClean et al. (2018), we measure gradient variance across 200 random initializations at each of 10 input points for the PauliZ observable, averaged over $q = 6$ qubits. Augmentation +Reupload has 22.122 against 9.568 for 1W+2D, $2.3\times$ the deployed gradient variance, while 2W+3D has the lowest of the five at 6.010; the latter is consistent with a more challenging landscape but does not diagnose its cause. These are trainability diagnostics only. At $q = 6$ every result here is classically simulated in a 64-dimensional state space.

O E5: THE FULL DEQUANTIZATION GRIDS

This appendix gives the complete arm accounting for the cached-map lane and the separate operational width audit. The two are labeled DQ-1 and DQ-2 in the protocol record and in the archive, and those labels are used here and in the reproducibility statement: DQ-1 is the 28-arm distillation judged by the three clauses of Section 4; DQ-2 is the finite-width q-scaling stop.

Section 4 reports that all 28 variant–student arms pass the ledger-declared builder-fit cost clause, 19 pass retention, and 4 pass all three arm-level clauses. Those counts are clause tallies over the same 28 arms, not a partition of them; the four matched-MLP arms, whose margins are the extreme ones, are reported by name below. Seeds are folded within dataset. The margin is student minus circuit downstream $R^2 \times 100$ under the fixed ridge readout and uses all 17 datasets. Retention is the mean of per-dataset student/circuit score ratios over 16 usable datasets; Bank32nh is excluded because its circuit score, 0.035, is below the ledger-declared 0.05 denominator floor. For the four arms that fire, retention is 99.56% for C0 with the degree-3 ridge, 100.11% for C0 with fitted RFF, 99.57% for C-edge with fitted RFF and 99.71% for C-GraphRZZ with fitted RFF, so the range the introduction quotes is 99.6–100.1%. Cost also uses all 17 datasets. Its decision compares mean student seconds with mean circuit seconds; the additional ratio summaries in Section 4 distinguish that ratio of means from the mean of per-dataset ratios. Table 8 sets out that clause accounting variant by variant.

Table 8: Cached-output conditional-distillation accounting under the ledger-declared builder-fit boundary. The arm-level analysis applies the three clauses marginally. A post-hoc exact Holm-28 sign-flip screening sensitivity retains the same four firing arms; its sign-symmetry assumption is unverified. RFF denotes random Fourier features.

Variant	Builder-fit cheaper	Mean retention $\geq 95\%$	Nominal CI lower end ≥ -1 pp	All three	Firing arm(s)
C0	7/7	5/7	2/7	2/7	Degree-3 ridge; fitted RFF
C-edge	7/7	5/7	1/7	1/7	Fitted RFF
C-GraphRZZ	7/7	4/7	1/7	1/7	Fitted RFF
C-reupload	7/7	5/7	0/7	0/7	—
Total	**28/28**	**19/28**	**4/28**	**4/28**	

[2] 3W+0D ends in a wide layer, so each final R_z angle is unobservable under the Z-basis diagnostic.

The ledger-declared readout-reproduction predictions hold at $R^2 \geq 0.95$ for C0 and C-edge, while the prediction that C-GraphRZZ would have the lowest best-student R^2 misses because C-reupload is lower (0.929 versus 0.974). The four firing-arm readout values are 0.955865 for C0/degree-3 ridge, 0.982040 for C0/fitted RFF, 0.977926 for C-edge/fitted RFF, and 0.974186 for C-GraphRZZ/fitted RFF. As a secondary comparison, every circuit clears the shared compressed-coordinate baseline by $+4.87$ to $+11.61$ pp. The per-dataset records range from 12–5 to 15–2 wins–losses, with no ties; all intervals for the mean circuit-minus-compressed-coordinate margin exclude zero, whereas only C-GraphRZZ's interval for the corresponding median margin reaches zero.

E5 students, splits, budgets, and cost fields. The full run is `run_id=e5_full`. A sealed 70/30 train/test split is used per dataset and seed; within training, a fixed disjoint 20% selection holdout is used. Circuit fitting is capped at 1,024 fit rows and 2,048 selection rows. All students map the q compressed coordinates to the circuit readout: degree-2 and degree-3 monomials followed by ridge ($\lambda = 10^{-6}$); fixed RFF with degree-2 width and median-bandwidth frequencies; fitted RFF with the same width and initialization, 200 Adam steps at learning rate 0.05, checkpointed every 10 steps on the inner holdout, then ridge; a parameter-matched one-hidden-layer ReLU MLP; a 64-unit wide MLP; and one histogram gradient-boosting regressor per readout coordinate with 100 iterations. Both MLPs use Adam, at most 600 iterations, early stopping, and a 15% validation split of training rows.

For DQ-1, ledger-declared circuit seconds equal circuit fitting plus the downstream ridge estimator's fit/test-chain seconds; student seconds equal student fitting plus the same downstream component. Those circuit seconds are 259.69 for C0, 231.91 for C-edge, 165.41 for C-GraphRZZ, and 569.68 for C-reupload. The four widest intervals in the grid are the matched-MLP arms at 15.51 to 17.63 pp, discussed by name below; the widest of the remaining 24 belongs to the C-reupload frozen-RFF arm at 7.47 pp, whose lower endpoint is -9.56 pp. Train-row teacher-output generation is timed and stored separately but excluded. Charging that recorded component to the student path while retaining circuit fitting is a construction-accounting sensitivity: for the four firing arms, the ratio-of-means magnitudes become approximately 117.2, 59.4, 56.1, and 50.3×, without changing their cheaper/not-cheaper indicators. Under a literal marginal-use sensitivity with circuit fitting sunk and teacher outputs cached, only C0 degree-3 ridge is cheaper for the observed batch; the three fitted-RFF circuit/student ratios are approximately 0.41, 0.40, and 0.36. If teacher-output generation is also charged to that one-batch student construction, none of the four is cheaper. These sensitivities answer different lifecycle questions and do not replace the ledger-declared primary accounting.

The arm with the extreme margins is the matched MLP. Its four means are -14.90 on C0, -15.81 on C-edge, -12.20 on C-reupload and -18.02 on C-GraphRZZ, at 0–17–0 everywhere except C-reupload, which is 1–16–0. A width-matched network fitted to reproduce the readout therefore does not come close on any variant. The budget the matching imposes is the reason: the ledger records this student as a dense $6 \to 2 \to 23$ network with 83 parameters against C0, C-edge and C-GraphRZZ, and $6 \to 3 \to 23$ with 113 against C-reupload's larger 124-parameter circuit. Because the 23 output biases alone consume much of a 76-parameter budget, the hidden layer is left at two or three units, which is what the wide student at $6 \to 64 \to 23$ and 1,943 parameters is run alongside to control for.

DQ-2 is the separate width sweep of run `e5_dq2_locked_20260813`. Its amended judgment cohort is SkillCraft, Bank32nh, Ailerons, Mercedes-Benz, CT Slice, and UJIIndoorLoc. The amendment followed feasibility failures, and the operational cohort-AND reduction was not fixed prospectively; the result is therefore non-confirmatory. At each dataset/seed/variant cell, the learner with highest held-out q=6 readout R^2 is carried unchanged across $q \in \{4, 6, 8, 10, 12\}$. Learner identity may vary across cells, and the same held-out split is used for selection and the reported trajectory.

All widths use 30 circuit steps, at most 128 circuit fit and selection rows, at most 512 student-fit rows, and at most 512 inference rows. The locked analysis validates the complete 960-row grid: 120 cell records and 840 student records, all successful, with no missing or duplicated entry. It first averages seeds within dataset, then requires every one of the six datasets to satisfy the cost-slope, q=12 cost, and minimum-R^2 clauses. No variant satisfies that cohort-AND rule. The main locked run used four concurrent workers; its wall-clock times are retained for provenance but are not used as isolated timing evidence. The cost clauses are evaluated separately by the sequential single-worker sensitivity. That post-hoc sensitivity runs one process and one numerical-library

thread. At every q it times one five-step full-batch circuit fit (five full-batch epochs), scales only that fit component by $30/5 = 6$ under a zero-intercept per-step model, and adds directly measured circuit inference and student fit/inference time. Reproduction scores are bound to the locked 30-step ledger; the five-step scores are diagnostics and never enter a verdict. The resulting cost clauses are an extrapolation sensitivity, not measured 30-step end-to-end timing or confirmatory evidence. The analysis validates 120 cells, 600 width points, and 600 short timing runs. In all 24 dataset–variant cells, the student total-cost slope is no larger than the circuit slope and its extrapolated total cost at q=12 is smaller. All sequential dataset-level clause outcomes match the primary concurrent analysis, including the locked 30-step R^2 clause; the cohort verdict remains false for C0, C-edge, C-reupload, and C-GraphRZZ because that reproduction clause fails on at least one dataset for each variant. The sequential result-record SHA-256 is `e24fda5ffa44fe3f4d43252987925241255104e39fa73d4417887396304cfd5a`.

The reproduction clause $R^2 \geq 0.90$ fails on one dataset for C0, two for C-edge, five for C-reupload, and one for C-GraphRZZ. C0 fails on Mercedes-Benz at 0.151 (q=10). C-edge fails on Ailerons at 0.732 (q=12) and Mercedes-Benz at 0.351 (q=10). C-GraphRZZ fails only on Mercedes-Benz, where the minimum is -0.474 at q=10. C-reupload fails on SkillCraft at 0.211, Bank32nh at 0.710, Ailerons at 0.816, and CT Slice at 0.892 (all q=12), and on Mercedes-Benz at 0.237 (q=10); UJI-IndoorLoc holds at 0.918 (q=10). Thus the locked reproduction evidence alone is already sufficient to prevent a DQ-2 stop for every variant, independently of the timing sensitivity.

The amended cohort uses six of the eight datasets initially eligible at $q = 12$. The two excluded curves are printed as sensitivity evidence rather than judged: Tecator and Residential. Their selected-student held-out R^2 at $q \in \{4, 6, 8, 10, 12\}$ is printed here to keep that registration. On Tecator: C0 0.989 and 0.957 with no further cell, C-edge 0.988, 0.952, 0.898, 0.875, 0.721, C-reupload 0.881, 0.716, 0.559, 0.359, 0.065, and C-GraphRZZ 0.980, 0.960, 0.840, 0.708, 0.599. On Residential: C0 0.990, 0.983, 0.962, 0.923, 0.935, C-edge 0.989, 0.989, 0.957, 0.935, 0.923, C-reupload 0.925, 0.876, 0.406, 0.768, 0.618, and C-GraphRZZ 0.988, 0.990, 0.972, 0.950, 0.945. The curves are partial where the registered width guard stopped the surrogate, which on Tecator left C0 with only q=4 and q=6; and at q=6 on Tecator the surrogate is row-limited (width 83 on 134 fit rows), so a falling R^2 there can be the surrogate running out of rows rather than expressiveness.

E5 reproduction scope. DQ-1 values are supported by the stored `e5_full` report and a reduction filtered to run `e5_full`. An unfiltered reduction can mix pilot rows; this is a supplementary-package reproduction-scope defect, not evidence that the displayed values are false. DQ-2 uses the fail-closed six-dataset reduction from run `e5_dq2_locked_20260813`; a general report's dataset-OR summary does not implement the cohort-AND rule and is not used for the manuscript verdict. The anonymous archive includes the compact DQ-1 report and a retrospective completeness validation of run `e5_full`: all 2,720 scoped rows are valid and successful, with no duplicates, missing cells, or unexpected family members. This is a present completeness check, not evidence that the protocol was prospectively enforced. The post-hoc exact sign-flip Holm-28 screening sensitivity and its output are also included; its sign-symmetry assumption has not been verified. That output is a one-sided sign-flip test of the dataset-level mean of student minus circuit plus one percentage point, Holm-corrected across all 28 contrasts at α=0.05, and its recorded status is a post-hoc deterministic completion of an underspecified preregistered instruction. It rejects on exactly four arms, which are exactly the four that fire on the three marginal clauses: C0 with the degree-3 ridge at p_{Holm}=0.0055, C0 with fitted RFF at 0.00021, C-edge with fitted RFF at 0.00060, and C-GraphRZZ with fitted RFF at 0.00021. The remaining 24 contrasts all sit at p_{Holm}=1. The agreement of the two selections is a property of this record, not a confirmation of either. The locked DQ-2 ledger, its analysis, source hashes, and compact outputs are included; the result remains operationally non-confirmatory; the A4 verdict is the cohort-AND reduction reported here, and the supplementary package records the permitted inference for each retained record.

P AUGMENTATION: ABLATIONS, STANDARDIZED CONTRASTS, AND POWER DIAGNOSTICS (E7)

The fixed-readout grid. Table 9 reports arm-level scores under both fixed readouts at T0 and, for Lasso, T1. T0 is the unstandardized per-arm design. T1 applies a fold-local standardizer to the complete per-arm design matrix $[X_{\text{raw}}, f(X)]$; it therefore also changes the raw columns and

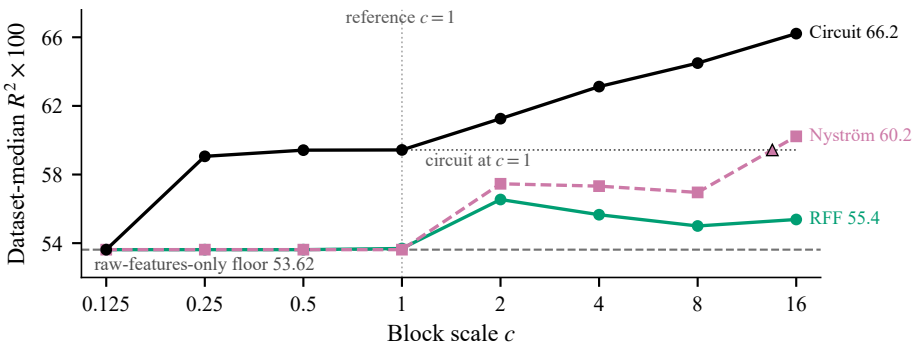

Figure 7: Block-scale sensitivity under a fixed Lasso readout (α=0.01). For seed 42, each point is the median $R^2 \times 100$ across the 17 regression datasets after multiplying only the mapped block by c and concatenating it with the unchanged raw block. The three arms are RFF, Nyström, and the circuit; no uncertainty intervals are available for this one-seed diagnostic, and connecting lines are visual guides. The dashed line is the raw-features-only median, 53.62, and the dotted lines mark c=1 and the circuit's value there. The triangle interpolates on the log scale between the Nyström points at c=8 and c=16 and is not an observation. OpenFE is omitted because its transform is not row-local and covers only 4 of 17 datasets.

changes the Original arm. The paired circuit-minus-arm columns report Lasso T0 only. Ridge T1 and the decisive Lasso T1 contrasts are recorded below rather than implied to be columns of this table.

Table 9: Fixed-readout performance, $R^2 \times 100$, as five-seed per-dataset means over the full 17-dataset descriptive cohort. T0 uses each arm's unstandardized full design; T1 applies a fold-local standardizer to its complete design matrix, including raw columns. The paired dataset-clustered circuit-minus-arm contrasts are Lasso T0 only. NPS is an approximate normal-power scale, $(1.96 + 0.84)$ times a standard-error proxy obtained from the bootstrap-CI width; it is not calibrated to the exact sign test or the registered joint mean/median/record rule and carries no formal verdict. No comparator is fitted to the circuit's readout; supervised transformers use the task label. Ridge T1 and selected Lasso T1 contrasts are reported below. OpenFE, available on only 4 of 17 datasets, is omitted from this row-local comparison and from the registered Holm reductions of Table 13. Splits are sealed; the T0 columns coincide with the comparison grid on 12 of 17 datasets, and on those 12 the deterministic baselines agree to within 4.7×10^{-5}. That tolerance describes the 12 coinciding datasets only: the means printed here are taken over all 17 and are not comparable to the comparison grid's. †Transformers absent from the five-comparator benchmark pool. PCA's T1 mean is dominated by one catastrophic cell ($\sim -10^{30}$) and is printed as 'diverges'.

	Ridge	Lasso		circuit − arm (Lasso, T0, n=17; CI of the mean)			
arm	T0	T0	T1	mean [CI_{95}]	median	W–L–T	NPS
circuit	**72.16**	57.85	**66.33**	—			
Original	8.16	3.11	15.78	+54.74 [+3.41, +146.1]	+3.91	13–2–2	102.0
PCA	18.55	12.20	diverges	+45.65 [+3.44, +119.1]	+3.91	13–3–1	82.67
Poly-2	−2,271,421.29	−2,201,518.56	−943,071.14	$+2.2\times10^6$ [−6.63, $+6.6\times10^6$]	+0.53	10–7–0	4.7×10^6
RFF	56.11	6.00	51.21	+51.86 [+1.64, +143.5]	+1.19	13–1–3	101.4
Nyström	63.60	9.45	57.66	+48.40 [+0.78, +135.0]	+1.16	9–5–3	95.90
Poly-3†	66.82	**61.37**	63.54	−3.52 [−8.80, +0.73]	+0.00	8–7–2	6.81
Trained RFF†	67.75	58.61	62.00	−0.76 [−3.99, +2.54]	−0.01	7–9–1	4.67
Budget-matched MLP†	61.45	54.70	58.35	+3.15 [−4.04, +14.27]	+0.11	10–5–2	13.08
Wide MLP†	67.77	59.72	63.20	−1.87 [−6.70, +2.81]	−0.35	5–11–1	6.79

Ridge under T1 and the Nyström sensitivity ladder. The full 17-dataset descriptive cohort is the complete record. The outcome-informed 15-dataset sensitivity removes SkillCraft and California Housing; the outcome-informed amended 13-dataset decision cohort additionally removes Kin8nm

and Space GA and is not untouched confirmation. On that amended decision cohort, Space GA's motivating old-grid residual of $+0.039$ pp had already become exactly zero under sealed remeasurement; its exclusion therefore does not supply independent support. On the amended decision cohort, Ridge-T1 circuit-minus-arm values in mean / median / record are Original $+5.46$/$+2.83$/12–1–0, Poly-2 $+0.68$/$+0.79$/9–4–0, RFF $+2.88$/$+1.39$/12–1–0, Nyström $+0.64$/$+0.53$/8–5–0, Poly-3 $+1.19$/$+0.54$/12–1–0, Trained RFF $+0.28$/$+0.62$/10–3–0, Budget-matched MLP $+0.43$/$+0.57$/9–4–0, and Wide MLP $+0.11$/$+0.11$/8–5–0. Against PCA, the Ridge-T1 circuit-minus-PCA mean diverges ($\sim +10^{32}$; median $+10.24$; 12–1–0). That divergence is a property of the contrast, in which PCA's own blow-up enters with a sign flip; it is a different quantity from the $\sim -10^{30}$ of Table 9, which is PCA's own Lasso-T1 score. None of the seven non-PCA decision comparisons clears the benchmark-derived three-part rule requiring both mean and median to exceed 2.0 pp with a favorable record. For reference, RFF has mean CI $[+1.17, +5.01]$ and an approximate normal-power scale of 2.75 pp; that proxy is not a formal verdict.

Table 10: Seven-arm non-PCA Lasso-T1 decision set. All contrasts are circuit minus arm on the outcome-informed amended 13-dataset decision cohort; this cohort is not untouched confirmation. The decision rule requires both mean and median to exceed 2.0 pp and a favorable record. Original is the sole raw-feature baseline; PCA and the post-hoc QKS pipeline are reported separately.

arm	cohort	mean	median	W–L–T	decision
Poly-2	13	$+0.75$	$\underline{+0.48}$	9–3–1	does not clear
RFF	13	$\mathbf{+2.13}$	$\mathbf{+1.02}$	10–2–1	does not clear
Nyström	13	$+0.14$	$+0.43$	7–5–1	does not clear
Poly-3	13	$\underline{+1.02}$	$+0.38$	9–3–1	does not clear
Trained RFF	13	-0.04	$+0.41$	8–4–1	does not clear
Budget-matched MLP	13	$+0.05$	$+0.25$	10–2–1	does not clear
Wide MLP	13	-0.04	$+0.02$	7–5–1	does not clear

Leave-one-out sensitivity of the raw-Original decision. On the 13-dataset decision cohort, removing any one of Yacht, Insurance, CPU Perf., Protein, Abalone, or UJIIndoorLoc lowers the Lasso-T1 circuit-minus-Original median from $+2.54$ to $+1.85$ pp (10–1–1) and the joint rule stops firing; the mean clause holds under every single deletion (lowest $+2.42$ pp without Yacht), and the rule fires on 57 of 78 leave-two-out pairs. On the untouched cohort only the removal of Forest Fires restores the mean clause (11 of 66 pairs fire). This is a post-lock analysis of already-scored ledgers (run `e7_locked_data_sensitivities_20260922`; no new cells) and changes no registered verdict.

Tuned-alpha sensitivities. The locked T2/T3 run contains 2,080/2,080 successful rows over the amended 13-dataset decision cohort, five seeds, eight arms, Ridge and Lasso heads, and two treatments. T2 selects alpha on inner folds without standardization; T3 applies fold-local standardization before the same selection. In both treatments, the representation has already been fitted once on the full outer-training split; only the downstream head is selected within inner folds. Thus T2/T3 estimate head-selection sensitivity conditional on a fitted representation, not fully nested representation tuning. Only Lasso enters the decision summary, with seeds averaged within dataset and T2 and T3 kept separate. None of the 14 contrasts in Table 11 clears the strict rule. The cohort and sensitivity were outcome-informed, so this complete grid tests robustness to alpha selection rather than providing untouched confirmation. The ledger SHA-256 is `0d41c2ddadf237e7542e61feb5d5a2385fa2afdcced16b4601530c961814eb4c`; the locked result-record SHA-256 is `f441a9316890b05dd25a7db49bd319b46beefcca7916aca89aa1fbd2eaa068a8`.

Table 11: Tuned-alpha Lasso sensitivities on the amended 13-dataset decision cohort. Entries are circuit-minus-comparator differences in $R^2 \times 100$ after averaging five seeds within dataset. Both treatments are head-only conditional sensitivities, not fully nested representation tuning. The strict decision requires mean and median above 2.0 pp and more wins than losses. Because T2 and T3 are different treatments, the bold and underline marks rank within each treatment block rather than down the whole column.

treatment	comparator	mean	median	W–L–T	decision
T2	Poly-2	-1.91	$\underline{+0.72}$	7–6–0	does not clear
T2	RFF	$\mathbf{+1.05}$	$\mathbf{+0.74}$	10–2–1	does not clear
T2	Nyström	-0.37	0.00	6–6–1	does not clear
T2	Poly-3	$+0.18$	$+0.42$	9–4–0	does not clear
T2	Trained RFF	$+0.06$	$+0.32$	9–4–0	does not clear
T2	Budget-matched MLP	$\underline{+0.22}$	$+0.41$	11–2–0	does not clear
T2	Wide MLP	$+0.03$	$+0.03$	7–6–0	does not clear
T3	Poly-2	$+1.36$	$\mathbf{+1.56}$	11–2–0	does not clear
T3	RFF	$\mathbf{+1.94}$	$+0.86$	10–3–0	does not clear
T3	Nyström	$+1.25$	$+0.53$	7–6–0	does not clear
T3	Poly-3	$+1.32$	$+0.22$	10–3–0	does not clear
T3	Trained RFF	$+1.60$	$\underline{+1.03}$	10–3–0	does not clear
T3	Budget-matched MLP	$\underline{+1.82}$	$+0.58$	10–3–0	does not clear
T3	Wide MLP	$+1.74$	$+0.05$	10–3–0	does not clear

Original, the sole raw-feature baseline, has a Lasso-T1 contrast of $+5.05$/$+2.54$/11–1–1 on this cohort. PCA is a fitted transform, but one catastrophic T1 cell makes its mean diverge; the $+8.64$ median and 11–1–1 record therefore do not define an interpretable joint-rule decision. Original lies outside the seven-transform decision family, so its tuned-head contrast is a separate outcome-informed sensitivity without familywise correction rather than part of the 14-contrast grid. In run `e7_t2t3_original_20260818`, Original is scored under the identical T2/T3 protocol (260/260 successful rows: 13 datasets, five seeds, both heads, both treatments), and the circuit side of each contrast is joined from the locked ledger with sealed-split hash equality verified on all 65 dataset–seed cells. Table 12 reports the Lasso decision coordinate. Boundary-grid alpha selections occur in 89/260 rows (34.2%), within the range observed in the locked grid, and are descriptive only. That pooled rate averages over four head–treatment coordinates whose rates differ by a factor of nearly three: Lasso-T2 33/65 (50.8%), Lasso-T3 24/65 (36.9%), Ridge-T2 20/65 (30.8%) and Ridge-T3 12/65 (18.5%). The decision coordinate is Lasso, so the rate that bears on Table 12 is the 50.8% of Lasso-T2 rather than the pooled figure. On ailerons under T2, the Original and circuit scores coincide exactly in all five seeds at the shared lower alpha bound, giving a zero contrast; under T3 the same dataset separates at $+11.57$ pp. The ledger SHA-256 is `3a070e3388551bec3d4017fa5fd4c30d4357c23a00dcae72c64fde0c2aee5996`; the result-record SHA-256 is `ce22091e713562b6a5ecf96245171bdfbef9ac48b398f4310bbd837382d580c3`.

Table 12: Tuned-alpha Lasso circuit-versus-Original sensitivity on the amended 13-dataset decision cohort, reported outside the seven-transform decision family. Entries are circuit-minus-Original differences in $R^2 \times 100$ after averaging five seeds within dataset, under the decision conventions of Table 11.

treatment	comparator	mean	median	W–L–T	decision
T2	Original	$\underline{+5.77}$	$\underline{+2.77}$	11–1–1	clears
T3	Original	$\mathbf{+6.74}$	$\mathbf{+3.99}$	12–1–0	clears

For Nyström under fixed Lasso, the T1 rungs are $+8.67$ with CI $[-0.02, +24.28]$, median $+0.63$, and 10–6–1 on the full 17-dataset descriptive cohort; $+1.37$ with CI $[-0.33, +3.37]$, median $+0.63$, and 9–5–1 on the outcome-informed 15-dataset sensitivity; and $+0.14$ with CI $[-0.95, +1.19]$, median $+0.43$, and 7–5–1 on the outcome-informed amended 13-dataset decision cohort. Their approximate normal-power scales are 17.37, 2.64, and 1.53 pp, respectively, and are not calibrated to the joint rule. The Lasso-T0 decision-cohort mean of $+0.72$ pp quoted in Section 5 is a different

quantity from the Ridge-T0 mean of $+0.72$ pp over the 16 datasets left after SkillCraft is removed; the two collide in rounding only and must not be quoted as one. Under T1, score coincidences with Original at tolerance 10^{-9} fall from 4/17 to 1/17 for Nyström and RFF and from 3/17 to 1/17 for the circuit; this does not imply identical predictions.

The registered Holm family. A nine-comparator Holm family was registered per head — Original, PCA, Poly-2, RFF, Nyström, Poly-3, Trained RFF, Budget-matched MLP and Wide MLP — and executed on the full 17-dataset descriptive cohort at T0, five seeds, 85 cells and 17 clusters per contrast, with the two-sided exact sign test on per-dataset differences as the per-contrast test. Table 13 reports all eighteen reductions. Three reject at α=.05: Ridge against Original (p_{Holm}=1.4×10^{-4}), Ridge against PCA (2.2×10^{-3}), and Lasso against RFF (1.6×10^{-2}). All three are contrasts against a raw or unsupervised-projection reference under a head whose own arm scores diverge, and each rejects on its win record while its mean stays unresolved: every one of the eighteen contrasts, the three rejections included, has a mean below its own approximate normal-power scale, so none of the point estimates is resolved by its interval. The correction therefore separates sign-consistency from effect size and does not license a magnitude claim. This family is also the exhibit for the fixed-Ridge circuit-minus-Nyström contrast that opens Section 5: $+8.56$ mean, -0.19 median, 7–10–0, with dataset-clustered mean CI $[-0.97, +25.44]$ and p_{Holm}=1.0, the discordance that leaves it inconclusive. The SkillCraft cell that drives that mean also carries the largest sealing shift: the Nyström score there moves from -33.00 unsealed to -74.82 sealed, which is what widens the SkillCraft circuit-minus-Nyström gap reported in Section 5. The manuscript's decision instrument is the joint mean/median/record point rule of Table 10, not this correction; the registered reductions are reported here for completeness and are not the decision procedure.

Table 13: The registered nine-comparator Holm family, both heads, on the full 17-dataset descriptive cohort at T0. Entries are circuit-minus-comparator differences in $R^2 \times 100$ after averaging five seeds within dataset. NPS is the approximate normal-power scale of Table 9; p is the two-sided exact sign test and p_{Holm} its Holm adjustment within the head's family of nine. Every contrast has a mean below its own NPS, including the three that reject. OpenFE is omitted for the reason given in Table 9. No entry carries emphasis marks: this correction is not the manuscript's decision procedure.

head	comparator	mean	median	W–L–T	NPS	p	p_{Holm}	reject
Ridge	Original	$+64.00$	$+5.83$	17–0–0	114.94	1.5×10^{-5}	1.4×10^{-4}	yes
Ridge	PCA	$+53.61$	$+5.21$	16–1–0	94.18	2.7×10^{-4}	2.2×10^{-3}	yes
Ridge	Poly-2	$+2.3\times10^{6}$	$+1.97$	13–4–0	4.8×10^{6}	0.049	0.343	no
Ridge	RFF	$+16.05$	$+0.64$	12–5–0	31.28	0.143	0.717	no
Ridge	Nyström	$+8.56$	-0.19	7–10–0	18.87	0.629	1.000	no
Ridge	Poly-3	$+5.34$	$+0.68$	12–5–0	6.67	0.143	0.717	no
Ridge	Trained RFF	$+4.41$	$+0.40$	11–6–0	8.55	0.332	0.997	no
Ridge	Budget-matched MLP	$+10.71$	$+0.75$	13–4–0	20.33	0.049	0.343	no
Ridge	Wide MLP	$+4.40$	$+0.22$	10–7–0	9.97	0.629	1.000	no
Lasso	Original	$+54.74$	$+3.91$	13–2–2	102.01	0.0074	0.059	no
Lasso	PCA	$+45.65$	$+3.91$	13–3–1	82.67	0.021	0.149	no
Lasso	Poly-2	$+2.2\times10^{6}$	$+0.53$	10–7–0	4.7×10^{6}	0.629	1.000	no
Lasso	RFF	$+51.86$	$+1.19$	13–1–3	101.42	0.0018	0.016	yes
Lasso	Nyström	$+48.40$	$+1.16$	9–5–3	95.90	0.424	1.000	no
Lasso	Poly-3	-3.52	0.00	8–7–2	6.81	1.000	1.000	no
Lasso	Trained RFF	-0.76	-0.01	7–9–1	4.67	0.804	1.000	no
Lasso	Budget-matched MLP	$+3.15$	$+0.11$	10–5–2	13.08	0.302	1.000	no
Lasso	Wide MLP	-1.87	-0.35	5–11–1	6.79	0.210	1.000	no

Untouched-cohort confirmation (one shot). The confirmation cohort contains 12 OpenML-CTR23 regression datasets selected by a mechanical rule locked before any score existed: of the 35 suite members, 12 were excluded by normalized-name identity with previously scored data, two as source-identity aliases, and two by the 195–58,310 row bound; among the survivors in ascending `data_id` order, four failed post-fetch gates (the row bound after cleaning, a non-regression target inference, or the six-feature minimum) and were replaced by the next surviving id, and the first 12 form the cohort (Table 15). Each dataset is version-pinned with its content hash recorded at fetch time. Run `e7_confirm_20260818` scores nine arms, both heads, T0 and T1, five seeds, and all 12 datasets: 2,160/2,160 successful rows, twelve dataset processes completing without error, and passing fixed-penalty, standardization, and sealed-split gates. The roster-argument extension is additive; on a 17-dataset-roster control cell its four coordinates are byte-identical to the archived grid

(split `434e4133c97b4b1c`). The single confirmatory coordinate, prespecified in the run lock, is the fixed Lasso-T1 circuit-minus-Original contrast under the unchanged joint rule, scored once and reported whichever way it falls.

The outcome is a confirmation failure: the joint rule is not confirmed. The mean clause fails at -21.95 pp while the median and record clauses pass ($+4.67$ pp, 10–1–1); the descriptive dataset-bootstrap 95% interval of the mean is $[-101.73, +25.46]$ pp (100,000 resamples). Table 16 lists the per-dataset contrasts. The mean is dominated by Forest Fires, where every mapped arm collapses under Lasso-T1 — RFF -554.6, circuit -464.4, Poly-2 -429.9, Poly-3 -376.2, Nyström -244.0 — while raw Original sits at -28.3: a collapse of wide mapped designs on one small, high-noise dataset expressed through the unbounded-below R^2 metric, not a circuit-specific failure. The mean clause's vulnerability to catastrophic cells is therefore a measured property of the rule. On Pumadyn-32nh the contrast is exactly zero. As descriptive context only, the circuit's Lasso-T1 median $R^2\times100$ on this cohort, 75.33, is the highest of the nine arms; Original's is 47.70.

Section 5.2 of the E7 preregistration declared in advance, and required to be reported beside the raw mean rather than in place of it, a robustness summary computed after clipping every arm's $R^2\times100$ below at 0. On this cohort that summary is $+14.38$ mean, $+4.67$ median, 10–1–1, which clears all three clauses. Clipping is applied per scored value before the per-dataset difference, as the preregistration words it; clipping the seed-averaged cell instead gives $+14.39$ and 10–0–2, so the choice does not move the verdict. This is a registered sensitivity, not a repair, and it does not displace the primary result: the confirmatory coordinate is scored on the raw metric, and on the raw metric the rule is not confirmed.

Post-hoc reductions of the confirmation grid. The run lock prespecified one coordinate only, the fixed Lasso-T1 circuit-minus-Original contrast. The same grid nevertheless scored the full nine-arm, two-head, two-treatment cross, so the remaining fixed-head circuit-versus-transform contrasts are reducible from it after the fact. Table 14 reports all 14, seven transforms under each readout. They are post-hoc reductions of a prespecified grid: the cohort and the scores are untouched, but the choice to read these coordinates was made after the confirmatory one had been scored, and they carry no confirmatory status and no multiplicity correction. One of the 14 clears the joint rule — Ridge-T1 against RFF, at $+7.79$ mean, $+3.33$ median and 8–4–0 — and it is not an artifact of the cohort's one catastrophic dataset: removing Forest Fires leaves $+5.99$/$+3.27$ at 7–4–0. The Lasso-T1 side, which is the readout the decision lane uses, clears 0 of 7 and so replicates the transform null the amended decision cohort records. Two Lasso-T1 contrasts pass the mean clause alone, against Poly-2 and RFF, and both fail the median clause, at $+0.05$ and $+0.83$ against the 2.0 pp threshold.

Table 14: All 14 fixed-head circuit-versus-transform contrasts on the untouched confirmation cohort under T1, in $R^2\times100$ after averaging five seeds within dataset over the 12 datasets. Positive favors the circuit. These are post-hoc reductions of a prespecified grid; only the Lasso-T1 circuit-versus-Original coordinate was prespecified, and it is reported above. The decision column applies the same joint rule as Table 10, requiring mean and median above 2.0 pp and a favorable record. No entry carries emphasis marks.

head	comparator	mean	median	W–L–T	decision
Lasso	Poly-2	$+4.37$	$+0.05$	6–5–1	does not clear
Lasso	RFF	$+15.42$	$+0.83$	6–5–1	does not clear
Lasso	Nyström	-13.49	$+0.08$	6–5–1	does not clear
Lasso	Poly-3	-0.18	$+0.47$	6–5–1	does not clear
Lasso	Trained RFF	-24.49	$+1.34$	8–3–1	does not clear
Lasso	Budget-matched MLP	-22.60	$+1.74$	10–1–1	does not clear
Lasso	Wide MLP	-20.96	$+0.48$	8–3–1	does not clear
Ridge	Poly-2	-15.39	$+0.02$	6–6–0	does not clear
Ridge	RFF	$+7.79$	$+3.33$	8–4–0	clears
Ridge	Nyström	-15.67	$+0.46$	8–4–0	does not clear
Ridge	Poly-3	-3.45	$+1.84$	8–4–0	does not clear
Ridge	Trained RFF	-24.47	$+2.30$	9–3–0	does not clear
Ridge	Budget-matched MLP	-21.74	$+2.15$	11–1–0	does not clear
Ridge	Wide MLP	-21.16	$+1.71$	9–3–0	does not clear

In the secondary, non-confirmatory paired control on the same cohort (run `e7_paired_control_confirm_20260818`; 60/60 cells; all four optimizer-entry hash

equalities pass), the trained-minus-zero-step contrast is -13.85 pp mean and $+1.24$ pp median with 10–1–1 and descriptive 95% CI $[-65.07, +17.93]$ pp, so its prespecified mean-based direction summary is not positive. Forest Fires again dominates (-272.35 pp; trained -464.35 versus zero-step -192.00), and on Pumadyn-32nh the trained and zero-step arms score identically. The selection-lock SHA-256 is `600b2944c84864fd2fe51c1d217bc1faedf5f039bc0c2ccb507194b43bb1a6e5`; the run-lock SHA-256 is `39552c4437e8ada3f87edb023cb96f836689ccff790c4b24d1b7752dc2410c1f`; the grid ledger SHA-256 is `cd362543685bb1ec9bab9a870a4173bb6df309d86e6d6390b59d95d98955005b`; the grid result-record SHA-256 is `9b672230ce4b8f96ad15a8290e8e3a499eb43dd7819a10127825675e4071a3c9`; the paired ledger SHA-256 is `945cd143335a6138b3193b5d40ba2124826d7dca18965d2b4f0733976d9ac0bd`.

Table 15: Untouched confirmation cohort: the first 12 surviving OpenML-CTR23 regression datasets in ascending `data_id` order, version-pinned at fetch time, after the locked mechanical exclusions.

dataset	`data_id`	version	n	d
Auction Verification	44958	7	2,043	7
Forest Fires	44962	9	517	12
Geo. Origin of Music	44965	7	1,059	116
QSAR Fish Toxicity	44970	7	908	6
Grid Stability	44973	7	10,000	12
SARCOS	44976	9	48,933	21
CPU Activity	44978	6	8,192	21
Diamonds	44979	8	53,940	9
Pumadyn-32nh	44981	2	8,192	32
Miami Housing	44983	2	13,932	15
CPS88 Wages	44984	2	28,155	6
Kings County	44989	2	21,613	21

Table 16: Untouched-cohort per-dataset contrasts in $R^2 \times 100$ after averaging five seeds within dataset. The primary column is the confirmatory fixed Lasso-T1 circuit-minus-Original contrast; the paired column is the secondary trained-minus-zero-step control. Neither column carries emphasis marks: the primary verdict is the prespecified joint rule, not a per-dataset comparison.

dataset	circuit − Original (T1)	trained − zero-step
Auction Verification	$+52.11$	$+21.40$
Forest Fires	-436.10	-272.35
Geo. Origin of Music	$+1.20$	$+1.15$
QSAR Fish Toxicity	$+3.11$	$+0.83$
Grid Stability	$+67.86$	$+62.39$
SARCOS	$+1.21$	$+0.36$
CPU Activity	$+14.43$	$+8.31$
Diamonds	$+4.17$	$+1.32$
Pumadyn-32nh	0.00	0.00
Miami Housing	$+16.90$	$+6.90$
CPS88 Wages	$+5.17$	$+0.09$
Kings County	$+6.51$	$+3.44$

Sealed block deletion. For the circuit under Lasso on the full 17-dataset, five-seed cohort, the block-on minus block-off contrast is -1.4319 mean, $+1.4994$ median, and 12–4–1 at T0; under T1 it is $+3.4138$, $+0.3338$, and 13–3–1. The mean changes sign and neither median is large, so the sealed control is unstable across treatments and does not establish robust benefit from retaining the raw block.

Untrained-pipeline and same-circuit controls. QKS is a distinct benchmark untrained random-circuit pipeline and is not a training-off version of the circuit arm. On the outcome-informed amended 13-dataset decision cohort and paired seeds 42 and 123, circuit minus QKS is $+4.48/+1.80$/10–1–2 for Lasso T0 (raw, unadjusted two-sided exact sign-test p=0.0117), $+4.50/+1.20$/11–1–1 for Lasso T1 (p=0.0063), and $+4.77/+2.14$/13–0–0 for Ridge T1 (p=0.0002), with the same raw test applying to all three values. No canonical adjusted QKS

result is retained, so no multiplicity-corrected inference is made. These are pipeline contrasts; QKS differs in circuit family, register sizing, number of circuits, Z-only readout, output width, and front end, and no cohort-level map-block scale comparison is reported. The same-circuit frozen-front-end control in Appendix E changes only the parameter tensor. On all 17 datasets and five seeds with Ridge, it gives $+1.00$ pp and 13–4 with Holm_5 $p{=}0.0895$, so it does not separate. This same-circuit control is Ridge-only and does not estimate the decisive Lasso training effect.

Q TASK SCREENING (E4): DEVELOPMENT, PREDECESSOR, AND GOVERNING RECORDS

The scored roster. The governing screen scores 34 arms per cell, and the arms outside the controlled family of Section 3 are named nowhere else in this paper, so they are specified here. Eleven are circuits. Four are the controlled variants C0, C-edge, C-reupload and C-GraphRZZ. Seven are decomposition arms built through the same registry. C-reupload triples the encoding, the wide layer and the deep layers together, so a null on it cannot say which change failed; *C-reupload-matched* and *C-deep-control* split that confound. C-reupload-matched re-uploads three times while holding C0's two-qubit gate count and depth; C-deep-control holds C-reupload's gate count and depth but uploads the data once. *C-GraphRZZ-2local* appends $3k$ two-local ZZ columns for the k selected edges, losing one column per chain-adjacent edge to duplication; *C-GraphRZZ-rankmatched* removes those duplicated columns outright; *C-GraphShadow* replaces the graph readout with all 1- and 2-local Pauli pairs read from one randomized-Pauli shot record, widening it to 63; and the two *-residual* arms, *C-GraphRZZ-residual* and *C-GraphShadow-residual*, select their sparse graph against the classical shell's out-of-fold training residual rather than against the unsupervised statistic. The residual arms therefore read y_{train} at graph selection, and a builder handed no labels raises rather than silently constructing the unsupervised sibling. Widths and gate-acting parameter counts at $q{=}6$ are 23/76 for C0, C-GraphRZZ, C-deep-control and C-GraphRZZ-residual; 23/96 for C-edge; 23/124 for C-reupload and C-reupload-matched; 38/76 and 35–38/76 for the two 2-local arms; and 63/76 for the two Shadow arms. The survivor rule runs separately in three width classes — *matched* at width 23, *wide* for the four wider circuits, and *concat* — with matched and wide candidates both judged against C0 and the same inner-selected width-23 bar, a comparison the reduction itself flags as not width-comparable for the wide class, and concat candidates judged against C0-concat and a bar inner-selected from the concatenated counterparts of the same six classical arms. The 46 drops are 22, 16 and 8 respectively, and the two survivors are both matched-class arms in the regression practical lane.

The six-member width-matched classical bar pool is `rff_learned` (23 columns, 161 fitted parameters), `rff_fixed` (23, unfitted frequencies), `mlp_matched` (23, 83), `mlp_wide` (23, 1,943), `poly2_trunc23` (degree-2 monomials truncated to 23) and `tensorsketch` (23). The bar is whichever of these six scores best in that cell, selected on inner training folds, not a fixed name. Four further arms are scored and reported but never judged, because their widths do not match: `poly2` (27), `poly3` (83), `z_floor` (the 6 compressed coordinates alone) and `raw_x` (the 6–454 raw features). `xgb_raw` is external: different features and a different estimator, reported as a ceiling and never as a control.

Those 22 arms are the unconcatenated field. The remaining 12 of the 34 are *-concat* variants — C0, C-GraphRZZ and C-GraphRZZ-residual, plus nine of the ten classical arms, all but `raw_x`, rescored with the raw features concatenated. They are judged inside their own *concat* class rather than against the width-23 bar, and are the subject of the separate concatenation analysis below.

Section 6 states that the regression and classification task blocks each have a fixed-head representation lane and an equal-budget, eight-trial practical lane. Crossing three circuit candidates with these two blocks and lanes gives twelve candidate–block–lane decisions, all of which fail the conjunctive requirement to clear C0 and the selected classical bar. The focal original development screen (run `e4v2_dev`) contains four datasets per task block, seed 0, 240 successful score rows plus eight audit rows, and no failures. Its score rows contain no inner-selection score. The deciding bar is therefore selected on held-out outcomes from the six-arm width-23 classical pool; it is the Wide MLP in every block–lane summary. This one-seed, post-outcome-selected development result is exploratory screening, not confirmatory evidence, and the later run's inner-selected choices are not part of this estimand.

The regression representation lane has C-edge clearing C0 by $+0.0093$ and losing the bar by -0.0147, C-reupload clears C0 by $+0.0185$ (4–0–0) and loses the bar by -0.0054, and C-GraphRZZ loses to both. The regression practical lane repeats those three patterns unchanged. On classification the representation lane has C-edge losing to C0 while C-reupload and C-GraphRZZ clear C0 and lose the bar; the practical lane has C-edge losing to both and the other two losing to C0 itself, by -0.99 and -0.64 pp. The width-unmatched strongest arm is the natural-width cubic basis, which beats all three circuits across both classification lanes by 2.1–4.9 pp. XGBoost on raw features (ceiling) is an external ceiling in both task blocks and lanes and sits about 0.10 sd above the best feature map in each regression lane. Neither the cubic nor raw-feature XGBoost is a width-matched decision bar, and neither enters the contrast families.

With four datasets in each family, the smallest attainable two-sided exact-sign p is 0.125, so no win count can reach $p < 0.05$. There are four separate Holm-8 families, one for each task-block–lane stratum. Every family contains the three candidate-versus-C0 and three candidate-versus-Wide-MLP contrasts, plus C0 versus Wide MLP and C0 versus the zero-feature floor; no member of any family rejects. Several approximate 95% resolution diagnostics are wider than the registered smallest effect size of interest (SESOI); these are not prospectively calibrated, and a null here cannot establish equivalence. This development analysis is separate from the later Holm-14 decomposition and the exploratory concatenation analysis.

The later concatenation analysis is outside registration and omits C-reupload. Against its selected bar, its four records are 1–3, 1–3, 1–2–1, and 2–2 for regression/representation, regression/practical, classification/representation, and classification/practical; the mean margin is negative in the first three. C-GraphRZZ-residual-concat clears both point hurdles only in classification/practical, at $+1.025$ pp versus C0 and $+0.289$ pp versus the inner-selected bar. This later, incomplete analysis is outside the four Holm-8 families, and its producing-source provenance gate is not closed, so the heterogeneous record receives no global verdict. The original screen likewise lacks its producing-source snapshot and supports only that no original candidate cleared both point hurdles, without directional, equivalence, replacement, or mechanism inference. The original and concatenation records are therefore artifact-only and not end-to-end regenerable from the archive. For both five-seed records below, the archive retains the producing driver and the ledger reduction but omits the roster container required to rerun the producer in a clean extraction.

The four-datasets-per-task predecessor, superseded as the governing record by the expansions below, runs the same development cohort under the retained driver (run `e4v2_rerun_20260818`, first 16 digits of the driver SHA-256 `265aed0e16a3cf2b`, seeds 42/123/456/789/2024). Amendment 6 was finalized before scoring and froze that driver and its default all-arm invocation, but its expected-volume line retained the obsolete 15-arm/1,200-score-row projection. The frozen driver expands that default to the 34-arm, two-lane roster established by the earlier budget-matched and concatenation additions. Execution produced 2,720 score rows plus 40 cell rows; the full-manifest check validates all 2,760 with zero failures. The frozen invocation, not the amendment's projection line, defines the executed roster. The deciding width-23 bar is selected on inner training folds per cell; it is the Wide MLP in all 20 regression cells and 15 of 20 classification cells (else truncated Poly2 or learned RFF). Every circuit is dropped in every task–lane–width stratum, and none of the 56 contrasts across the four Holm-14 families rejects. Directionally, C-reupload minus C0 remains positive at $+0.0195$ sd mean ($+0.0171$ median, 3–1–0) on regression representation and $+1.07$ pp ($+0.89$, 3–1–0) on classification representation, consistent with the original one-seed $+0.0185$; the budget-matched C-reupload-matched is $+0.0181$ sd ($+0.0178$, 4–0–0, exact-sign p=0.125); and every circuit loses to the inner-selected bar on regression representation (0–4–0). That budget-matched figure is superseded by the governing run, which also scores its sibling decomposition arm beside it: there C-reupload-matched minus C0 is $+0.0126$ sd ($+0.0047$ median, 10–3–0) on regression representation and $+0.0144$ ($+0.0041$, 10–3–0) on regression practical, and C-deep-control minus C0 is $+0.0022$ ($+0.0004$, 8–5–0) and $+0.0030$ ($+0.0001$, 7–6–0) in the same two lanes. On classification the same four contrasts are $+0.42$ pp ($+0.33$, 8–5–0) and $+0.41$ ($+0.07$, 9–4–0) for C-reupload-matched, and $+0.28$ ($+0.23$, 8–5–0) and $+0.03$ ($+0.27$, 9–4–0) for C-deep-control. The four-dataset exact-sign floor of p=0.125 is unchanged in that record, so it sharpens provenance, seeding, and bar selection, not inferential reach. Ledger SHA-256 `a57d1a1033485806...`; report SHA-256 `f498c6a6e160eba6...`.

That floor is a property of cohort size alone, $2/2^n$, and was computable before any score existed, so the eight-per-task predecessor, run `e4v2_expand16_20260824`, doubled the cohort

to eight datasets per task under a preregistration written before scoring (SHA-256 `08816b31-9caaf1f6...`), with the datasets fixed by name order from the roster and the prior expectation recorded. Its eight retained datasets were re-scored and reproduce the earlier run on all 2,720 cells in split hash, score, and prediction hash, so the two runs are comparable. The floor falls to p=0.0078 and two contrasts attain it, both the same comparison in the two regression lanes: C-GraphRZZ-residual minus the width-matched wide MLP at -0.0234 sd on representation and -0.0223 sd on practical, each 0–8–0 with p_{Holm}=0.109. No contrast rejects in that record, 48 arms drop, and none survives; the bar is the Wide MLP on 32 of 40 regression representation selections and 33 of 40 practical ones. Holm-14 correction over each family puts the smallest attainable adjusted value at $14 \times 0.0078125 = 0.109375$, so no outcome at that cohort size could have produced a corrected rejection: that expansion converted an arithmetically vacuous record into a multiplicity-blocked one rather than into a powered test. The first cohort size at which a corrected rejection becomes attainable is ten datasets per task, where $14 \times 2/2^{10} = 0.02734375$; nine still falls short at 0.0546875. Ledger SHA-256 `8261e155a1b9d508...`; report SHA-256 `2856619990aa5ba1...`.

The governing record: thirteen datasets per task. The governing run, `e4v2_expand26_20260827`, takes the cohort to thirteen datasets per task under Amendment 8 (SHA-256 `2918e4432d83f3c2...`), written before any cell of the expansion was scored, and Section 6 reports it. The cohort is derived by the predecessor's own rule with one word changed: the four development-cohort members of each lane are retained, and the prefix rule that Amendment 7 applied to the first four byte-ordered roster candidates is extended to the first nine. Because it is a prefix rule over a fixed, already-sorted list, extending it cannot disturb any member already chosen; positions five through nine add cpu_performance, ctslice, diabetes, insurance and kin8nm to regression and har, isolet, jasmine, madelon and mfeat_factors to classification. Byte-wise order decides the head of one list, since `M` precedes `a`, so Mercedes_Benz precedes ab_regression and any case-folded re-derivation would select a different set. The size, thirteen, was chosen in knowledge of the predecessor's 0–8–0 contrasts, because it is the smallest cohort whose corrected rejection survives one dataset going the other way; which datasets enter was outcome-blind, and the two decisions are recorded separately rather than blended. The run scored 130 units across 34 arms and five seeds with zero failures and an 8,970-row ledger. Its sixteen retained datasets reproduce the predecessor on all 5,440 re-scored cells in split hash, score, and prediction hash, twice the 2,720 that record could check, and the registered family gate passes: the family size is 14 with identical membership in all four task–lane strata. The inner-selected bar remains predominantly the Wide MLP, on 49 of 65 regression representation cells, 54 of 65 regression practical, 44 of 65 classification representation and 43 of 65 classification practical, with learned RFF, the matched MLP and truncated Poly2 taking the remainder; the measured bar therefore did not change identity under the expansion, which the family gate exists to confirm. Ledger SHA-256 `78d87cae1b484d62...`; report SHA-256 `eaf4bce36f42e7aa...`.

At thirteen datasets the exact-sign floor is $2/2^{13} = 0.000244$ and the first Holm step admits $14 \times 0.000244 = 0.003418$, so a corrected rejection is reachable, and it survives one broken sweep: a 13–0 record gives $p_{\text{Holm}} = 0.003418$ and a 12–1 record 0.047852, while 11–2 returns 0.314453 and is retained. That 11–2 row was written down before scoring precisely so it could not afterwards be narrated as marginal. **No contrast attains the floor.** The smallest raw p observed is 0.003418, the 12–1 value, and seven contrasts sit there; all seven reject at $p_{\text{Holm}} = 0.047852$ and no other contrast rejects. Effective n matters here because the test counts non-zero paired differences: two contrasts recorded one exact tie each and therefore ran at $n = 12$ — C-GraphRZZ-residual minus C0 and C0 minus `z_floor`, both in classification, where balanced accuracy over a finite test set is discrete — and neither is among the seven. Every rejecting contrast carries $n_{\text{nonzero}} = 13$, as the registered gate requires.

The seven corrected rejections. They divide three ways. Four are a circuit below the width-matched wide MLP: C-GraphRZZ-residual at -0.0246 sd on regression representation, -0.0211 sd on regression practical and -1.78 pp on classification practical, and C-GraphRZZ at -1.30 pp on classification representation. One is a circuit below the base circuit, C-GraphRZZ-residual minus C0 at -0.0133 sd on regression practical. Two are C0 above the trivial floor, $+0.127$ sd on both regression lanes. All seven are 12–1–0 records at $p_{\text{raw}} = 0.003418$ and $p_{\text{Holm}} = 0.047852$. The universal that holds over all seven is narrower than "classical beats circuit", and is stated at that width: no rejection places a circuit above a width-matched classical arm. The four wide-MLP

contrasts run classical-over-circuit, which is the direction the amendment registered as its expectation; the residual-minus-C0 contrast is circuit-over-circuit and belongs to neither direction; and the two C0-minus-z_floor contrasts do place a circuit above a classical arm, but z_floor is the six compressed coordinates alone, a width-6 arm outside the width-23 matched pool, so beating it is not a width-matched comparison and licenses nothing about the bar. The sign test and the interval estimate do not agree everywhere, and the disagreement is reported rather than resolved: five of the seven carry dataset-clustered 95% intervals excluding zero, but the two C-GraphRZZ-residual-minus-mlp_wide regression contrasts do not — $[-0.0513, +0.0008]$ on representation and $[-0.0497, +0.0085]$ on practical — so the reduction files them as inconclusive on the interval instrument while the corrected sign test rejects. Nothing is claimed from the mean for those two beyond the sign.

The two survivors, and why they are not a falsification. Both are reuploading arms in the matched class of the regression practical lane, and both clear the registered point hurdle, which asks only that a candidate beat C0 and the inner-selected bar on the mean. C-reupload beats C0 by $+0.0159$ sd and the bar by $+0.0032$; C-reupload-matched by $+0.0144$ and $+0.0017$. The records run the other way: against the bar they are 2–11–0 and 3–10–0, so each wins on the average of thirteen datasets while losing eleven and ten of them. In the Holm families the corresponding contrasts against mlp_wide are $+0.0081$ and $+0.0066$ sd on 4–9–0 records at $p_{\text{Holm}} = 1.0$, with intervals $[-0.0157, +0.0472]$ and $[-0.0201, +0.0474]$ that cover zero and escape the ± 0.02 sd SESOI; the recorded verdict on both is INCONCLUSIVE. Neither beats the strongest arm at any width: against raw_x they are -0.0054 and -0.0069 sd on 6–7–0. The registered falsification clause requires a candidate to beat both references, exceed the block SESOI, and carry an interval excluding zero; none of the three margins exceeds the SESOI and the family intervals cover zero, so the clause is not met and no advantage is claimed. What can be said is narrower and is said as such: the two arms' nominal, uncorrected intervals against C0 do exclude zero, $[+0.0029, +0.0306]$ and $[+0.0045, +0.0271]$ sd, at means well inside the SESOI, and the de-confounded arm shows the same direction on the representation lane ($+0.0126$ sd, 10–3–0). Amendment 8 registered in advance that a surviving circuit would be a reportable surprise rather than a protocol failure, and it is reported here on that footing.

An unequal head-selection budget on four datasets. The practical lane's eight-trial head-selection cap did not bind equally across the expanded cohort, and the run's consistency checks flag it. Scanning all 4,420 practical score rows, the cap bound on exactly one arm — the external xgb_raw ceiling — on 16 of the 130 cells: ctslice completed 4–5 of 8 trials at every seed, har and isolet 1 of 8 at every seed, and mfeat_factors 7 of 8 at seed 42. No circuit, no width-matched classical arm, and no reference entering any contrast family was capped, so the decisive comparisons and the survivor comparisons are unaffected. The representation lane is untuned throughout, so the ceiling figures Section 6 quotes from it carry none of this exposure either. The disclosure is recorded because an unequal tuning budget would weaken comparability had it fallen on a judged arm, and because the affected quantity, the boosting ceiling, is if anything understated on those four datasets.

The run's consistency checks raise one further flag. isolet, one of the thirteen classification datasets and one of the four just named, is a declared/inferred task conflict: the roster manifest declares it classification while the inferred label is regression, and all five of its cells carry the task-conflict flag. The executed run takes the task from the manifest's explicit declaration, which is the correct label for a 26-class problem, so no score, contrast, or verdict is affected; Amendment 8 §9 added the dataset in knowledge of the conflict and said so in advance. It is named here because the same consistency checks that report the tuning cap report this, and a run described as having zero failures should not be read as having raised nothing.

The registered survival reference, and what the run executed. The reference the survival rule is drawn against changed after registration; the chronology is as follows. The E4-V2 preregistration, §6, registered a named reference: a candidate circuit survives only if it beats *both* C0 *and* rff_learned, and its §6 contrast set is eight members per block, three candidates crossed with {C0, rff_learned} plus C0 − rff_learned and C0 − z_floor. Its amendment replaced that named reference with a measured one — the highest-scoring width-matched classical arm in that block and lane, drawn from a pool fixed in advance as rff_learned, rff_fixed, mlp_matched, mlp_wide, poly2_trunc23 and tensorsketch — on the ground that

rff_learned was not the strongest width-matched arm in either block: in the four-dataset pilot mlp_wide beat it by 0.100 sd on regression and 7.57 pp on classification, and beat every circuit in the same table. That amendment states on its own first line that it was written after the pilot results were seen. Amendment 2 then moved the selection from the outer test split to inner training folds, since choosing the bar on held-out scores is test-set selection applied to the baseline. Amendment 7 restates the executed rule and fixes the family structure at four families of 14. The executed run matches those amendments and not §6: the deciding bar is inner-selected per cell from the six-arm pool, and the Holm family's classical slot is mlp_wide, with no rff_learned member anywhere in any of the four families. rff_learned remains a scored pool member and is printed beside the bar, but it does not define it.

The 14 family members are identical in all four task–lane strata, which the report's own family lists confirm; enumerated once, they are C-edge, C-reupload, C-GraphRZZ, C-reupload-matched, C-deep-control and C-GraphRZZ-residual, each against C0 and against mlp_wide, plus C0 − mlp_wide and C0 − z_floor.

Per-contrast intervals in the governing run. The analysis emits an interval verdict with every contrast: a dataset-clustered 95% CI, whether that CI excludes zero, whether the contrast falls within the registered SESOI for its block (0.02 sd regression, 1.0 pp classification), and whether the cohort resolves that SESOI at all. Table 17 counts them over the 56 contrast records. Across the four strata, 24 of 56 CIs exclude zero, 11 fall within the block SESOI, and 29 are powered for it. The predecessor's statement that the power repair was real but uneven has been re-derived here rather than carried over, as Amendment 8 required, and at thirteen datasets it points the other way. Resolution is $\text{mde}_{95} = 1.96\,\text{sd}/\sqrt{n}$, so it improves with cohort size only when the spread of paired differences does not grow faster. It did grow in regression: the median minimum detectable effect on the representation lane rises from 0.0100 to 0.0270 sd, above that block's 0.02 sd SESOI, and the powered count falls from 12 to 6 in each regression lane. In classification the representation-lane median fell, from 1.5383 to 0.9618 pp, below that block's 1.0 pp SESOI, and the powered count rises from 3 to 7 on representation and from 2 to 10 on practical. The practical lane moves the same way in both tasks, to 0.0301 sd and 0.8815 pp. The equivalence verdicts move with it and are no longer a regression phenomenon: 7 of the 11 now fall in the two classification lanes, against 1 of 15 in the predecessor. The five heavy regression datasets the amendment added are what widened that spread, and the sign test, which reads only signs, gained resolution over the same cohort where the interval estimate lost it. These are nominal, uncorrected, per-contrast intervals. No familywise equivalence procedure was registered, and none is run: the 11 individually SESOI-bounded intervals of this run do not compose into a familywise equivalence claim, and the manuscript makes none. The direction of the equivalence finding was itself registered. The E4-V2 preregistration, §9, written before scoring, predicts no advantage in either block: no candidate circuit beats both C0 and rff_learned at once, and the primary contrast's dataset-clustered 95% CI is expected to fall inside the block SESOI. Its falsification clause is that a candidate beats both references, exceeds the block SESOI, and has a CI excluding zero. Reporting the interval verdicts is therefore reporting against a registered prediction, not a post-hoc equivalence argument. The predecessor's two floor-attaining contrasts, C-GraphRZZ-residual minus mlp_wide in the two regression lanes, were resolved well enough there to call the loss a difference rather than a null: minimum detectable effects of 0.0127 and 0.0124 sd against intervals $[-0.0355, -0.0119]$ and $[-0.0342, -0.0111]$. In the governing run the same two contrasts reject on the corrected sign test but no longer clear that bar, at 0.0269 and 0.0299 sd against intervals $[-0.0513, +0.0008]$ and $[-0.0497, +0.0085]$ which cover zero. They are the clearest case of the divergence above: the record hardens from 0–8–0 to 1–12–0 while the interval widens past the SESOI, so the rejection is carried by the sign test alone and the magnitude is not resolved.

Sources of the screen's three summary quantities. Section 6 quotes three figures from this run. All three are derived below from the governing report; the superseded four-dataset record gives different values and is not their source. The raw-feature XGBoost ceiling sits $+0.1191$ sd above the best width-matched circuit in the regression representation lane, where that circuit is C-reupload, and $+7.8696$ pp above it on classification, where it is C-reupload-matched; the practical lane gives $+0.1153$ sd and $+7.7261$ pp. The width-matched qualifier still does work in the classification figure: reading against the widest circuit scored, the 63-column C-GraphShadow, the same ceiling is $+7.4634$ pp. That gap widened from the eight-per-task record's $+4.2972$ pp, and both sides move: on the thirteen-dataset cohort the boosting mean is 2.02 pp higher while the three arms that deter-

Table 17: Interval verdicts recorded by the governing run for the 56 contrasts of the four Holm-14 families, counted by task–lane stratum. The 12 survivor-block records are excluded. "Excludes zero" and "within SESOI" are read from each contrast's dataset-clustered 95% CI against the block SESOI of 0.02 sd (regression) or 1.0 pp (classification); "powered" is whether the cohort resolves that SESOI. All three are nominal and uncorrected, and no familywise equivalence procedure was registered.

task	lane	contrasts	excludes zero	within SESOI	powered
Regression	representation	14	5	2	6
Regression	practical	14	6	2	6
Classification	representation	14	7	4	7
Classification	practical	14	6	3	10
All		56	24	11	29

mine the ceiling figures — C-edge, C-reupload-matched and `mlp_wide` — fall by 1.55 to 1.80 pp. Across all thirteen width-23 arms the change is not uniform and is not described by that band: on the representation lane it spans $+0.13$ pp (`tensorsketch`, the one arm that rises) to -3.27 pp (`mlp_matched`, which falls furthest, not `mlp_wide` at -1.79). These are cohort means over different dataset sets rather than within-dataset changes, so the widening is a property of which datasets are in the cohort and not a movement of any arm on a fixed one; the identity of the best width-matched circuit moves with it, from C-edge to C-reupload-matched. XGBoost reads different features under a different estimator and enters no contrast family. C0 loses to its own inner-selected class bar 4–9–0 in both regression lanes, at -0.0139 sd on representation and -0.0127 on practical, where the bar is inner-selected per cell and therefore differs from the `mlp_wide` rows of Table 18 — the baseline circuit is beaten by the classical arm the candidates are also measured against, in the lane where the feature map is the only difference. The natural-width cubic exceeds the best width-matched circuit on classification only, as it did in the predecessor: $+1.2474$ pp on the representation lane and $+1.2283$ on the practical lane, both against C-reupload-matched, but -0.0251 and -0.0169 sd on the two regression lanes, where the circuit is ahead of it. Like XGBoost, `poly3` carries 83 columns against the shared 23 and is judged in no family.

The full circuit-versus-width-matched-MLP contrast set. Table 18 gives all 28 contrasts against `mlp_wide` — the seven width-23 circuits, C0 among them, in each of the four task–lane strata — so the one contrast Section 6 prints can be read against the rest of its set rather than on its own.

R E6: THE SPLIT-ENCODING HYBRID, EVERY JUDGED CELL

The retained secondary diagnostic has the hybrid clearing its in-slot classical control in two of eight judgment cells, with neither clearance surviving the second cohort; Table 19 prints all twelve cells that were judged — the eight target-y cells the body counts and the four residual-target cells that ran beside them — so each clearance can be read against its own detectable effect and no failing cell is left out of view. A clearance requires the mean, the median and the win–loss record to agree in sign against that control; where they do not, the table names the dataset carrying the mean. Regression scores are error-based and pre-oriented by the report so that positive favors the hybrid; classification is balanced accuracy in pp. The 4×2 routing cross (140 cells over the four pilot datasets) ran to completion and is included in the supplementary material.

S E2 AND E8: PREDICTIONS, FAMILY MARGINS, TRANSFER PANELS

The eight registered predictions. The immutable identities below follow the eight rows of the E2 preregistration, §5. Each receives exactly one status; the status counts are descriptive per-family or conditional rule outcomes, not an experiment-wide multiplicity-controlled success rate. All configuration-cluster intervals and Holm results below are conditional on the registered grid over five fixed, reused dataset geometries; they do not generalize across datasets.

P1 (missed): Pairwise product, low b: C-GraphRZZ was registered to beat C0 while Degree-2 should not lose because the target lies in span. The primary clause missed at -1.64 pp, me-

Table 18: All 28 circuit-minus-`mlp_wide` contrasts in the governing run, over thirteen datasets per task. Regression is in target-standard-deviation units and classification in balanced-accuracy percentage points; positive favors the circuit. `mlp_wide` is the width-23 wide MLP, which is also the arm the inner selection picks for the bar in most cells. No entry carries emphasis marks.

task / lane	circuit	mean	median	W–L–T
Regression, repr.	C0	−0.0093	−0.0036	4–9–0
	C-edge	−0.0069	−0.0074	4–9–0
	C-reupload	+0.0043	−0.0063	3–10–0
	C-GraphRZZ	−0.0272	−0.0189	2–11–0
	C-reupload-matched	+0.0033	−0.0033	4–9–0
	C-deep-control	−0.0072	−0.0090	2–11–0
	C-GraphRZZ-residual	−0.0246	−0.0201	1–12–0
Regression, pract.	C0	−0.0078	−0.0035	4–9–0
	C-edge	−0.0031	−0.0072	5–8–0
	C-reupload	+0.0081	−0.0063	4–9–0
	C-GraphRZZ	−0.0233	−0.0179	2–11–0
	C-reupload-matched	+0.0066	−0.0024	4–9–0
	C-deep-control	−0.0048	−0.0089	4–9–0
	C-GraphRZZ-residual	−0.0211	−0.0202	1–12–0
Classif., repr.	C0	−1.14	−0.40	4–9–0
	C-edge	−0.75	−0.25	4–9–0
	C-reupload	−0.86	−0.38	4–9–0
	C-GraphRZZ	−1.30	−0.94	1–12–0
	C-reupload-matched	−0.72	−0.60	3–10–0
	C-deep-control	−0.86	−0.73	4–9–0
	C-GraphRZZ-residual	−2.12	−0.93	2–11–0
Classif., pract.	C0	−0.80	−0.58	6–7–0
	C-edge	−0.55	−0.63	4–9–0
	C-reupload	−0.72	−0.42	5–8–0
	C-GraphRZZ	−1.06	−0.80	3–10–0
	C-reupload-matched	−0.39	−0.47	5–8–0
	C-deep-control	−0.77	−0.70	4–9–0
	C-GraphRZZ-residual	−1.78	−1.09	1–12–0

Table 19: E6, all judged hybrid-minus-in-slot-control cells. Positive values favor the hybrid. Regression effects use target-standard-deviation units (sd), and classification effects use balanced-accuracy percentage points (pp). The development and headroom cohorts each contain eight datasets in total, four per task block, and overlap on two; every row below is therefore a four-dataset record, which is why no exact sign p falls below the $n{=}4$ floor of 0.125. MDE_{95} is the half-width of the dataset-clustered 95% interval; the registered smallest effect sizes of interest are 0.02 sd for regression and 1.0 pp for classification. The "+" column identifies which of mean, median, and win–loss–tie record are positive; clearance requires all three.

block	lane	target	cohort	mean	median	W–L–T	sign p	MDE_{95}	unit	+	carried by
regr.	repr	y	development	+0.0313	+0.0072	4–0–0	0.125	0.0494	sd	all three	–
regr.	repr	y	headroom	+0.0082	−0.0002	2–2–0	1.000	0.0216	sd	mean only	Residential
regr.	repr	residual	headroom	−0.0315	−0.0144	1–3–0	0.625	0.0610	sd	none	Space GA
regr.	practical	y	development	+0.0162	+0.0062	4–0–0	0.125	0.0213	sd	all three	–
regr.	practical	y	headroom	+0.0065	−0.0036	1–3–0	0.625	0.0225	sd	mean only	Residential
regr.	practical	residual	headroom	−0.0316	−0.0145	1–3–0	0.625	0.0617	sd	none	Space GA
class.	repr	y	development	+0.3858	−0.2386	1–2–1	1.000	1.7744	pp	mean only	Sonar
class.	repr	y	headroom	+0.4456	−0.0558	1–2–1	1.000	1.7615	pp	mean only	Sonar
class.	repr	residual	headroom	−0.2600	+0.0558	2–1–1	1.000	1.7114	pp	not the mean	Mfeat-Pixel
class.	practical	y	development	+0.5859	−0.2185	1–2–1	1.000	2.8162	pp	mean only	Vowel
class.	practical	y	headroom	−0.6000	−0.5682	1–2–1	1.000	1.1368	pp	none	Sonar
class.	practical	residual	headroom	−1.3615	−0.8621	1–2–1	1.000	1.8128	pp	none	Mfeat-Pixel

dian −0.59, 6–36; the Degree-2 comparison was sign-split at −0.05, median +0.65, 27–15. **P2 (untestable):** Pairwise product, high b: the circuit was to beat Degree-2 and learned-frequency RFF only if β reached b. The gate was unmet ($|\beta| = 1.75$ versus 18.20; 0% of cells reached the registered 80%); the descriptive contrasts were +18.29 (37–5) and −3.81 (12–30). **P3 (fired):** Two-scale sinusoid: C-reupload beat C0 by +2.12 pp (median +1.78, 60–24, Holm 0.0356). **P4 (fired):** On the same family, learned-frequency RFF beat both C-reupload (+1.38, 49–35) and C0 (+3.50, 56–28); this separately registered record says the observed axis is not quantum-specific. **P5 (fired):** The grouped Linear, Friedman-1, and Degree-2 controls were level or classical-favored: 0–42 three

times on Linear, 0–42 three times on Degree-2, and ± 1.0 pp equivalence on Friedman-1, with the negative-control condition not triggered. **P6 (quarantined; protocol-bug condition triggered and unresolved):** Clifford teacher showed no advantage over the post-outcome named best classical arm (-0.13, 15–27), but the Degree-2 contrasts triggered the registered protocol-bug condition at C0 $+6.81$ (41–1, Holm 0.0078), C-GraphRZZ $+6.28$ (41–1), and C-reupload $+6.22$ (39–3); this is an advantage over one arm, not the classical set, and its cause is unresolved. **P7 (missed):** Sign parity predicted C-GraphRZZ first among circuits and XGBoost first overall. C-GraphRZZ was lowest of the four circuits, although the XGBoost on the latent (ceiling) was first of 11 at $+4.03$ (100–25). **P8 (missed):** Quantum teacher predicted a C-GraphRZZ win, but the named contrasts were -0.42 (11–31) against C0 and -0.51 (2–40) against C-edge.

The original contrast tests are Holm-adjusted only within their respective family sets; there is no Holm family over these eight statuses. The later named-best and per-cell-envelope summaries are post-outcome descriptive comparisons, not selection-adjusted inference; the width-unmatched XGBoost result is a descriptive ceiling.

Family margins. No family has a resolved designated-circuit lead over its named best classical arm at the registered margin. Seven named-best point estimates are negative, from -0.09 to -5.81; Sign parity alone is positive but inconclusive. In the registration order, each record gives the designated circuit against the family's named best classical arm, then mean, median, win–loss, configuration-cluster CI of the mean, descriptive interval half-width, and the win–loss against the per-cell classical envelope: linear target, C0 against the degree-2 basis, -0.18, -0.03, 0–42, $[-0.37, -0.05]$, 0.16, envelope 0–42; Friedman-1, C0 against the degree-3 basis, -0.09, -0.01, 19–23, $[-0.17, +0.01]$, 0.09, envelope 16–26; degree-2 target, C0 against the degree-2 basis, -1.18, -0.55, 0–42, $[-1.99, -0.43]$, 0.78, envelope 0–42; Clifford teacher, C-GraphRZZ against the single-layer MLP, -0.13, -0.16, 15–27, $[-0.53, +0.31]$, 0.42, envelope 10–32; quantum teacher, C-GraphRZZ against the learned-frequency RFF, -0.23, -0.27, 10–32, $[-0.42, -0.03]$, 0.19, envelope 4–38; two-scale sinusoid, C-reupload against the single-layer MLP, -2.54, -0.90, 19–65, $[-3.84, -1.19]$, 1.32, envelope 16–68; pairwise product, C-GraphRZZ against the single-layer MLP, -5.81, -3.46, 8–76, $[-7.87, -3.89]$, 1.99, envelope 6–78. The eighth, sign parity, C-GraphRZZ against the degree-3 basis, has the sole positive named-best point estimate and is inconclusive: $+1.51$, $+0.78$, 69–57, $[-0.53, +3.68]$, half-width 2.10, envelope 56–70. Its interval uses 27 $(\text{variant}, \text{noise}, N)$ configuration clusters over 126 paired cells. Against Degree-2 and learned-frequency RFF separately, its designated circuit has means $+8.29$ pp (Holm 7×10^{-5}) and $+3.40$ pp (Holm 1.5×10^{-5}), but it is -1.18 pp against C0 and -0.83 pp against the per-cell classical envelope; these comparator-conditional contrasts do not override the inconclusive named-best result.

Under the unadjusted descriptive ± 1.0 pp convention, the selected-comparator summaries classify the linear target, Friedman-1, the Clifford teacher and the quantum teacher as level, and the degree-2 target, two-scale sinusoid and pairwise product as classical leads. The registration tags the linear target, Friedman-1, the degree-2 target and the Clifford teacher as controls, and the quantum teacher's answer as planted. The best classical arm is chosen after seeing scores, so these summaries are not selection-adjusted. A five-dataset exact sign-flip analysis has floor $p = 2/2^5 = 0.0625$; this resolution limits cross-dataset inference from the repeatedly used geometries. Two further contrasts were registered as descriptive, outside the Holm family and carrying no verdict: C-GraphRZZ over TensorSketch at $+12.3/+8.3$/461–43–0 and over the fixed-frequency RFF at $+8.5/+4.6$/461–43–0, 504 cells each, seed 0, diagnostic rows excluded. A family names a target-generating process; the pairwise-product family includes pairwise product (low b) and pairwise product (high b) settings. A best classical arm names a feature map.

E8 transfer panels. Raw dimension is an exploratory k=4-specific proxy from five fixed E2 geometries, distinct from the pooled k=2, 3, 4 Sign-parity statistics. The E7 outcome scores already existed. Under the asserted chronology, the 12/5 split by raw input dimension, T1 statistic, and decision rule were fixed before E8 margin inspection or aggregation, not before E7 cells were scored. This is a semi-confirmatory secondary analysis of reused scores rather than independent confirmation. Table 20 prints every one of the 17 targets with the dimension that assigned it and both heads' margins, and marks the two overlapping E2 high-dimensional inputs removed by the registered 15-dataset sensitivity. Margins are circuit minus per-cell classical envelope, $R^2 \times 100$; the

low-dimension group was registered to favor the circuit. Ridge and Lasso are reported separately; the registration specified no prospective cross-head joint success or multiplicity rule. The registered 2.0 pp line is an analysis convention borrowed from repeatability, not a practical-significance threshold.

Table 20: E8 per-dataset T1 margins, ordered by raw input dimension. Margins are circuit minus the per-cell classical envelope in $R^2 \times 100$ percentage points, so positive favors the circuit. The heads are fixed Lasso and Ridge readouts. "Group" applies the registered raw-dimension split at 100 (12 low, 5 high). "In 15-panel" marks inclusion in the registered sensitivity that removes the two E2-overlapping high-dimensional datasets, Mercedes-Benz and CT Slice.

dataset	raw dim	group	ℓ_1 T1	ridge T1	in 15-panel
Insurance	6	low	-0.69	-0.37	yes
Space GA	6	low	-7.09	-1.46	yes
Yacht	6	low	$+0.01$	$+0.05$	yes
CPU Perf.	7	low	-6.66	-6.70	yes
Abalone	8	low	-1.49	-1.86	yes
California	8	low	-2.19	-1.30	yes
Kin8nm	8	low	-2.94	$+0.55$	yes
Protein	9	low	-4.33	-4.57	yes
Diabetes	10	low	-13.03	-10.59	yes
SkillCraft	18	low	$+27.49$	$+9.43$	yes
Bank32nh	32	low	-0.47	-0.96	yes
Ailerons	40	low	$+0.00$	-0.76	yes
Residential	107	high	-0.61	-0.67	yes
Tecator	124	high	-0.16	-0.12	yes
Mercedes-Benz	356	high	-0.02	-0.28	no
CT Slice	384	high	-2.50	-2.28	no
UJIIndoorLoc	454	high	$+0.45$	$+0.48$	yes

Table 21 aggregates those per-dataset margins into the registered panel and prints the same panel again under three descriptive shakes of the registered coordinate, so the reader can check that the between-group difference runs against the registered direction everywhere the panel was formed and not only at the coordinate the body quotes. Each registered row is one head crossed with one treatment, carrying the mean and median margin and the win–loss–tie record for the 12 low-dimension and the 5 high-dimension datasets, then the between-group median difference, the U statistic and the one-sided p of the registered exact test. Only T1 is a decision coordinate; T0 and the other shakes are descriptive. The low-dimension records on the two T1 rows — the side registered to favor the circuit — are 3–9–0 on the ridge head and 2–9–1 on the Lasso head. A shake row re-forms that panel with one coordinate changed and prints only what the re-formed panel decides: seed-average first averages seeds before the envelope is formed rather than after, envelope without the Original arm drops that arm from the classical envelope, and rank correlation correlates the margin with raw dimension across all 17 datasets instead of splitting them into two groups, which is why it prints no median difference and why its p is a two-sided correlation p rather than the one-sided rank-sum p of every other row. The registered coordinate is single (the E8 preregistration, §5), so no shake row enters the verdict of Section 6.

The registered 15-dataset sensitivity removes Mercedes-Benz and CT Slice, the two overlapping E2 high-dimensional inputs, leaving 12 low and 3 high targets. At T1 the registered median/rank direction remains reversed and the rule is not supported: Ridge has median difference -1.015 pp, mean difference -1.442, $U = 9$, and one-sided $p = 0.9099$; Lasso has -1.688, -0.845, $U = 8$, and $p = 0.9319$. These values do not establish a general absence of sign-interaction transfer; the proxy fails the preregistered support rule and gives adverse-direction point estimates on the reused E7 cohort even after removing the two overlaps.

Two perturbations close the panel: swapping the group labels moves the one-sided p to 0.19 (ridge) and 0.12 (ℓ_1), and in 2,000 fixed-seed (seed 7) Monte Carlo 12/5 relabelings the observed one-sided p-values have estimated ranks 80.2 and 87.4, defined as 100 times the fraction of relabeled p-values strictly below the observed value. This is a descriptive diagnostic, not predictive accuracy. The exact test's reachable floor, $p = 0.00016$, shows only that discreteness does not block $p < 0.05$; it does not establish adequate power.

Table 21: E8 17-dataset panels (12 low / 5 high). "Head" is the fixed Ridge or Lasso readout; T0 is the unstandardized design and T1 is fold-local standardization. At the registered coordinate, per-seed circuit-minus-classical- envelope margins are formed before seed averaging with Original included. Margins and the low-minus-high median difference are in $R^2 \times 100$ percentage points; positive favors the registered low-dimension direction. Win–loss–tie records compare the circuit with the envelope. "Shake" rows change exactly one coordinate—seed averaging order, Original's inclusion, or the grouped statistic (rank correlation)—and are descriptive, outside the verdict, with only the resulting between-group statistic shown.

head	treat	coord.	low mean/med	low W–L–T	high mean/med	high W–L–T	med. diff	U	p
Ridge	T0	registered	$-0.484/-1.550$	1–11–0	$-1.049/-0.128$	2–3–0	-1.422	18	0.9027
Ridge	T1	registered	$-1.545/-1.130$	3–9–0	$-0.574/-0.280$	1–4–0	-0.850	21	0.8358
Lasso	T0	registered	$-13.364/-8.345$	0–12–0	$-0.362/-0.141$	2–3–0	-8.204	4	0.9989
Lasso	T1	registered	$-0.950/-1.844$	2–9–1	$-0.566/-0.155$	1–4–0	-1.688	18	0.9027
Lasso	T1	*shake:* seed-average first					-1.567	–	0.9682
Lasso	T1	*shake:* envelope without the Original arm					-1.688	–	0.9027
Lasso	T1	*shake:* rank correlation					–	–	0.0987
Ridge	T1	*shake:* seed-average first					-0.804	–	0.8087
Ridge	T1	*shake:* envelope without the Original arm					-0.850	–	0.8358
Ridge	T1	*shake:* rank correlation					–	–	0.5020

The companion verification script reports the sibling exact-match counts, margins, label-swap test, and fixed-seed relabeling diagnostics. It performs no enforcement, so a discrepancy would be reported but would not make the check fail closed.

T E3: ROUTED TWO-QUBIT GATE COUNTS

The routed-count diagnostic records the re-uploading variant as the family's largest tested circuit on this resource axis; the ordering holds at every qubit count transpiled. Per stacked circuit block the routed two-qubit gate counts at $q = 4, 6, 8, 10$ are 180, 465, 849 and 1,287 for C-reupload; 54, 155, 275 and 429 for C0 and, identically, for C-edge; and 24, 43, 77 and 108 for C-GraphRZZ. Counts are taken after transpilation to a line coupling map (basis $\{\mathrm{R}_Z, \sqrt{X}, X, \mathrm{CNOT}\}$, optimization level 1, fixed transpiler seed), verified by the pinned re-transcription script included in the supplementary material. This is a hardware-resource axis, distinct from the circuit-versus-surrogate cost axis of Section 4; no circuit here ran on hardware, and these counts support no practical-cost claim.

U E1: SPAN AND FOURIER FITS ON THE FULL SWEEP

The controlled E1 fits are teacher-dependent readout-reconstruction diagnostics, not teacher-free task models. They are held out on the full 17-dataset sweep, roster-enforced. At each cell, a ridge map from the basis to the circuit readout is fitted on the training split and scored on held-out readouts as $R^2 \times 100$. The declared E1 convention uses the natural basis width and weights readout columns by their within-cell variance; seeds are folded within dataset before dataset-clustered means and intervals. On the degree-2 span the values are 91.58 $[88.36, 94.42]$ for C0, 90.60 $[87.19, 93.59]$ for C-edge, 82.93 $[78.54, 87.05]$ for C-reupload and 90.00 $[86.84, 92.60]$ for C-GraphRZZ; on the learned-frequency basis, in the same order, 98.23 $[97.28, 99.02]$, 97.83 $[96.76, 98.70]$, 91.59 $[88.81, 94.16]$ and 97.47 $[95.96, 98.59]$. These four learned-frequency levels quantify readout reconstruction and do not establish task substitution, function equality, or broader practical equivalence. The C-edge-minus-C0 quadratic contrast is -0.988 pp mean, -0.797 pp median, 3–14–0, with interval $[-1.663, -0.431]$ pp and Holm p=0.0127; its learned-frequency contrast is -0.401 pp, 1–16–0, with interval $[-0.682, -0.050]$ pp and Holm p=0.00110. That learned-frequency interval excludes zero and lies wholly inside the registered ± 1 pp span-score margin, so difference and margin equivalence hold simultaneously without implying function equality, task substitution, teacher-free replacement, or a mechanism.

The full adjusted family is $\{\text{C-edge}, \text{C-reupload}, \text{C-GraphRZZ}\} \times \{\text{quadratic}, \text{learned-frequency RFF}\}$ against C0. In that order within each circuit, the mean differences (pp), records, and Holm-adjusted p-values are: C-edge, -0.988/3–14–0/0.0127 and -0.401/1–16–0/0.00110; C-reupload, -8.658/0–17–0/0.0000916 and -6.641/0–17–0/0.0000916;

and C-GraphRZZ, -1.581/1–16–0/0.00110 and -0.758/2–15–0/0.00470. All six exact sign tests reject after Holm correction.

The deployed-reference calculation is a different pipeline and compression. Its four descriptive conventions are width-truncated/uniform 94.965%, width-truncated/variance-weighted 96.813%, natural-width/uniform 95.686%, and natural-width/variance-weighted 97.280% $[96.08, 98.21]$. The last convention was adopted after the sweep began and is post-hoc rather than prospectively confirmatory. It must not be used as a numerical bridge to controlled E1's C0 quadratic value of 91.58% $[88.36, 94.42]$. The two 91.6 entries within E1 are also different quantities that collide only in rounding: C0's degree-2 span is 91.5846 and C-reupload's learned-frequency fit is 91.5856.