

Entanglement entropy is not a faithful explanation: auditing a financial quantum feature map

Jean Marie Vianney Sindayigaya

Abstract—Entanglement entropy is an attractive candidate for explaining quantum machine-learning models: it is intrinsic to the encoding, cheap to compute, and requires no measurement. We test whether it is a faithful feature attribution for a financial-transaction quantum feature map, under a protocol predeclared down to its thresholds, its comparison family, and its posture on failure. Two questions are separated by construction. First, does the operating point selected for accuracy leave usable attribution signal? Second, does that signal correspond to what actually drives the model's decisions? On the Kaggle Credit-Card Fraud dataset over ten repeated case-control splits, per-feature entropy dispersion is 71% lower at the accuracy-selected entanglement strength than at the midpoint of the grid, and the difference is significant after Holm–Bonferroni correction; however, a lower-strength regime with $3.4\times$ the dispersion is available at no statistically detectable accuracy cost. Faithfulness fails decisively. Entropy rankings are significantly *anti*-correlated with model-aware occlusion attributions (median Spearman -0.299), and on a synthetic task whose informative features are known by construction and which the quantum model demonstrably learns, occlusion recovers ground truth ($+0.237$) while entropy does not (-0.025), both significant after correction. A model-randomisation check identifies the cause exactly: per-feature entropy is invariant to the trained model, with maximum change of exactly zero across all ten seeds, so it cannot distinguish a correctly trained model from one fitted to shuffled labels. We conclude that entanglement entropy is a genuine and inexpensive diagnostic of the encoding but not an explanation of the model, and we report this refutation as predeclared. Balanced-cohort metrics are benchmark values, not deployment-prevalence estimates.

Keywords—Quantum Machine Learning; Explainable AI; Feature Attribution; Entanglement Entropy; Quantum Kernel Methods; Financial Fraud Detection; Negative Results

I. INTRODUCTION

EXPLAINING a quantum machine-learning model is usually approached by importing a classical attribution method — Shapley values, local surrogates, permutation importance — and paying its cost in circuit executions [1]–[4]. An appealing alternative is to read the explanation off the quantum state itself. A feature map entangles the qubits carrying different features; the entanglement entropy of a feature's reduced state is intrinsic to the encoding, requires no measurement, and costs a single state preparation for all features at once. If that quantity tracked what the model uses,

quantum models would come with an explanation essentially for free.

This paper tests whether it does, for the financial-transaction feature map FT-FM and its tunable-entanglement refinement introduced earlier in this series [5], [6]. The question matters beyond convenience: credit and fraud decisioning are Annex III high-risk uses under the EU Artificial Intelligence Act, whose Articles 13 and 86 require transparency and explanation of individual automated decisions [7], [8], and the adequacy of an explanation is a technical property before it is a legal one.

We separate two questions that are usually conflated. First, does the operating point chosen for *accuracy* leave usable attribution signal in the encoding? The refinement of [6] exposes an entanglement-strength parameter β , and its validation protocol selected the maximum $\beta = 1$ in every seed; near-maximal entanglement is precisely the regime where per-feature entropies risk collapsing toward a common value. Second, and independently, does entropy correspond to what drives decisions? A quantity can vary richly across transactions and still explain nothing.

Both questions are answered by a protocol fixed in advance — thresholds, comparison family, claim rule, and the commitment to publish refutations. Following the framing decision recorded before any confirmatory run, the paper is written around the questions rather than a presupposed answer.

Our contributions are as follows. 1) We define entanglement-based attribution for feature maps whose features may occupy several qubits, with a decision-level occlusion attribution as the faithfulness target (Section III). 2) We predeclare a six-comparison family under one Holm–Bonferroni correction, a saturation threshold fixed before any run, and a synthetic track whose informative features are known by construction (Section IV). 3) We report that entropy dispersion collapses by 71% at the accuracy-selected operating point, yet a $3.4\times$ more dispersed regime costs no detectable accuracy (Section V). 4) We report that entanglement entropy is nonetheless *not* faithful: it is significantly *anti*-correlated with model-aware attribution, significantly fails to recover known ground truth where a model-aware baseline succeeds, and is invariant to the trained model with maximum change exactly zero — the mechanism that explains the other two results (Sections V–VI).

Section II reviews attribution and its evaluation, Section III defines the method, Section IV the predeclared protocol,

Author is with Warsaw University of Technology, Poland (e-mail: jean_marie_vianney.sindayigaya.dokt@pw.edu.pl).



Section V the results, Section VI discusses and bounds them, and Section VII concludes.

II. BACKGROUND AND RELATED WORK

A. Attribution and what makes one trustworthy

Post-hoc feature attribution assigns each input a share of responsibility for a prediction, either axiomatically [3] or through a local surrogate [4]; the field’s vocabulary and taxonomy are surveyed in [9]. The evaluation literature is more instructive than the method literature for our purposes. Explanations must be tested rather than admired [10]; many saliency methods fail elementary sanity checks, most decisively the requirement that an explanation of a model must *change when the model changes* [11]; stability under input perturbation is a separate and often unmet requirement [12]; and permutation-based importance evaluates a model off the data manifold, so it is a flawed reference rather than a gold standard [13]. Rudin argues that in high-stakes settings post-hoc explanation should be abandoned in favour of intrinsically interpretable models [14] — an objection this paper engages directly, since the quantity we test is intrinsic to the encoding yet, as we show, still not an explanation of the model.

B. Explainability for quantum models

Existing quantum XAI largely transplants classical machinery. Shapley values have been computed over circuit *gates*, attributing outcomes to components of the circuit rather than to input features [1]; local surrogates have been applied to quantum neural networks [2]; and applied studies in security domains report attribution for quantum classifiers [15]. All inherit the sampling cost of the classical method they import. The alternative examined here is to use a quantity the quantum state already provides. For a pure global state, the von Neumann entropy of a subsystem’s reduced density matrix is the operationally meaningful measure of its entanglement with the remainder [16], [17], so a feature’s entropy is a natural candidate for “how much this feature participates in the encoding”.

C. The encoding under study

FT-FM allocates encoding axes by feature type and couples qubits on a domain-informed graph [5]; FT-FM-lite scales the entangling interactions by $\beta \in [0, 1]$, leaving the per-feature encoding axes untouched [6]. Its fidelity kernel is consumed by a support-vector classifier [18], [19]. Two prior results frame this paper. Validation selected $\beta = 1$ in all ten seeds on the Kaggle Credit-Card Fraud dataset [6], [20] — the maximum of the grid. And highly expressive quantum kernels concentrate, losing input-dependent structure [21], which raises the possibility that the accuracy-selected regime is also the least informative one to inspect. A companion study of a gated hybrid detector built on the same encoding found that its quantum branch contributed no net corrections on this dataset [22], which makes the question of what that encoding actually represents more pressing rather than less. Whether that is so, and whether it matters for faithfulness, are the questions of Sections V and VI.

III. ENTANGLEMENT-BASED ATTRIBUTION

A. Encoding-level quantities

Let $|\psi(x)\rangle = U_{\text{FT}}(x; \beta)|0\rangle^{\otimes n}$ be the encoded state and let feature j own the qubit set Q_j . Features need not occupy one qubit each: a binary-encoded categorical with K values occupies $\lceil \log_2 K \rceil$ qubits, so attribution is defined on *qubit groups*. Writing $\rho_{Q_j} = \text{Tr}_{\overline{Q_j}}|\psi\rangle\langle\psi|$, the per-feature entanglement entropy is

$$S_j(x) = \frac{-\text{Tr} \rho_{Q_j}(x) \log_2 \rho_{Q_j}(x)}{|Q_j|}, \quad (1)$$

normalised to bits per qubit so that features of different width remain comparable within a ranking; for a single-qubit feature the denominator is one and $S_j \in [0, 1]$. For a pure global state this is the operationally meaningful entanglement measure of the subsystem [16]. The pairwise quantum mutual information $I_{jk} = S_j^{\text{raw}} + S_k^{\text{raw}} - S_{jk}^{\text{raw}}$ is reported in raw bits, being a joint quantity with no natural denominator.

Two properties of Eq. (1) are immediate and consequential. It depends only on the feature map and the input, never on the trained classifier or the labels; and all n feature attributions are read from a single state preparation, so the encoding-level explanation costs one circuit execution regardless of feature count.

B. Decision-level attribution as the faithfulness target

Since S_j is model-free, it cannot by itself be an explanation of a model’s decision. We therefore adopt a model-aware attribution as the target against which faithfulness is measured. With f the deployed decision function and b a predeclared baseline vector,

$$\Delta_j(x) = f(x) - f(x^{(j \rightarrow b_j)}), \quad (2)$$

the change in the decision when feature j is replaced by its baseline. The baseline is the per-feature training-cohort median, fixed before any confirmatory run. Occlusion costs one kernel row per feature, i.e. n rows against the fitting cohort, and unlike permutation importance it never evaluates the model on permuted — and therefore off-manifold — combinations of correlated features [13].

C. Saturation

A feature is *saturated* at a given β when the standard deviation of S_j across the held-out cohort falls below $\varepsilon = 10^{-3}$: its entropy is then effectively the same for every transaction, so it cannot distinguish them and can carry no attribution signal whatever its mean value. The threshold is two orders of magnitude above double-precision eigenvalue noise for a one-qubit reduced state and was fixed in the specification before any confirmatory run. We also report *dispersion*, the mean across-cohort standard deviation of S_j over features, as a continuous measure of how much signal the encoding retains.

D. What a faithful attribution would have to satisfy

Three requirements are testable here and are tested in Section V. A faithful attribution should (i) agree with a model-aware attribution of the same decisions, (ii) recover informative features on a task where they are known by construction, and (iii) change when the model changes [11]. Requirement (iii) is the sanity check that no model-free quantity can pass, and we predeclare its failure rather than discovering it: a quantity independent of the trained model cannot distinguish that model from one fitted to shuffled labels. We measure it anyway, because the magnitude of that invariance is the mechanism that explains whether (i) and (ii) can hold.

IV. PREDECLARED EXPERIMENTAL METHODOLOGY

Every element below — operating points, thresholds, comparison family, claim rule, and the posture on failure — was fixed in a written specification before any confirmatory evaluation.

A. Pilot disclosure

An exploratory pilot on one seed’s training cohort motivated the hypotheses: it indicated that per-feature entropies at $\beta = 1$ were concentrated near their maximum with little across-sample variation. We disclose it rather than presenting the hypotheses as though formed in ignorance. No pilot number appears in any result of Section V; the confirmatory protocol re-tests each claim on held-out cohorts across all ten seeds at thresholds fixed in advance.

B. Data, cohorts, and tracks

The real-data track reuses the frozen protocol of the previous papers in this series on Kaggle Credit-Card Fraud [6], [20]: ten seeds, stratified 60/20/20 raw splits before any fitting, training-only Fisher selection of five PCA components alongside Time and Amount, and deterministic 1:1 case-control cohorts of 240/80/80 rows. All reported F1 values are therefore case-control benchmark metrics, not deployment-prevalence estimates; the raw positive rate is $492/284,807 \approx 0.17\%$. Attributions are computed on test cohorts only, after all models are fit, using exact statevector simulation [23].

The synthetic track exists because Kaggle CCF has no known ground-truth feature importance. Its generator labels a transaction anomalous when the amount, the time of day, or the velocity counter crosses a threshold, so exactly three of its seven features are informative by construction and the remaining four — a merchant category and three relational components — are noise. This mask is a property of the generator, not a choice made for this paper.

Amendment (dated before the amended run; both runs reported). The initial run drew synthetic cohorts by stratified split of a 200-sample pool. With a $\approx 7\%$ positive rate the models received about nine positive training examples and underfitted (test F1 0.49–0.55), leaving the ground-truth comparisons *inconclusive rather than negative*: an attribution cannot be shown to recover informative features through a model that has not learned them. The track was amended to

draw disjoint 1:1 case-control cohorts of 240/80 from a 4000-sample pool, identical in construction to the real-data track, and a classical RBF-SVM reference was added per seed so that residual underfitting could be attributed to the encoding rather than to task difficulty. The hypotheses were unchanged in form, and every real-data value is bit-identical across the two runs.

C. Operating points and hypotheses

The grid is $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$. Two points are named in advance: the *deployed* $\beta = 1$, inherited from the previous paper’s unanimous validation selection, and the *informative* $\beta^* = 0.5$, fixed a priori at the midpoint rather than selected per seed, which removes a researcher degree of freedom.

The confirmatory family contains exactly six two-sided Wilcoxon signed-rank comparisons [24] on paired per-seed statistics under a single Holm–Bonferroni correction at familywise $\alpha = 0.05$ [25]. Directionality lives in the claim rule, not the test: a *win* requires Holm significance *and* a median difference in the predicted direction, so a significant result in the wrong direction is reported as a loss with equal prominence.

- H2a Entropy dispersion is higher at β^* than at $\beta = 1$.
- H2b Entropy dispersion is higher at $\beta = 0.25$ than at $\beta = 1$.
- H3 Entropy and occlusion rankings agree (Spearman > 0) at β^* on Kaggle CCF.
- H5a Entropy recovers synthetic ground truth at β^* .
- H5b Occlusion recovers synthetic ground truth at β^* .
- H6 Test F1 at β^* is lower than at $\beta = 1$.

Two further checks are non-inferential and sit outside the family. H1 asks whether a majority of features are saturated at $\beta = 1$, reported with per-seed counts. H4 measures the change in each attribution when the fitted model’s labels are randomised; entropy is predeclared to be exactly invariant and occlusion to change.

Effective pair counts, dropped-zero counts, and exact attainable p -floors are reported per comparison, so a reader can verify that each test was decidable rather than inferring decidability from a nominal $n = 10$.

D. Refutation posture and provenance

We predeclare that failures are published as refutations and not reframed, that non-significance is never reported as equivalence, and that the paper is written around its questions rather than a presupposed answer. Every number derives from archived artifacts carrying their repository commit, a clean working-tree flag, pinned package versions, and the input file’s SHA-256 digest.

V. RESULTS

All values derive from archived runs of ten seeds under the canonical Python 3.11 environment with a clean working tree. Balanced cohorts make every F1 a case-control benchmark metric.

TABLE I
ENTANGLEMENT-STRENGTH SWEEP ON KAGGLE CCF OVER TEN SEEDS. DISPERSION IS THE MEAN ACROSS-COHORT STANDARD DEVIATION OF PER-FEATURE ENTROPY. VALUES ARE CASE-CONTROL BENCHMARK METRICS.

β	Mean entropy	Dispersion	Test F1
0.00	0.0000	0.0000	0.864
0.25	0.2387	0.0801	0.877
0.50	0.6302	0.0856	0.879
0.75	0.9071	0.0409	0.890
1.00	0.9763	0.0248	0.891

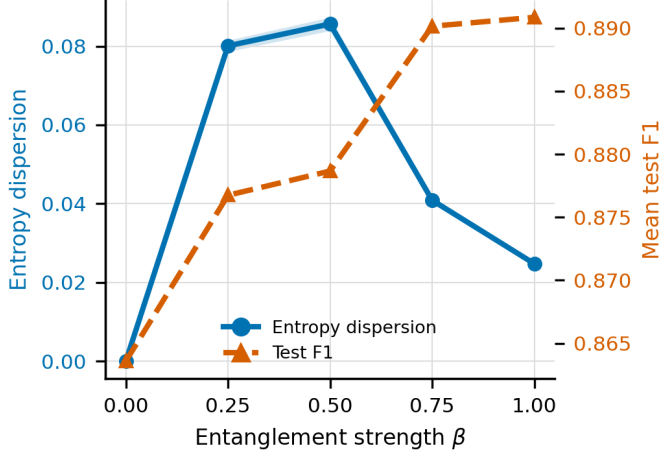


Fig. 1. Entropy dispersion (left axis) and test F1 (right axis) against entanglement strength. Dispersion peaks at $\beta = 0.5$ and collapses toward the accuracy-selected $\beta = 1$; shading is the 95% bootstrap interval of the mean.

A. Saturation and dispersion

H1 is refuted. At the predeclared $\varepsilon = 10^{-3}$, exactly *one* of seven features is saturated at $\beta = 1$ — in all ten seeds, without exception — rather than the predeclared majority. The threshold was fixed before any confirmatory run and has not been moved.

What holds instead is compression rather than saturation, and it is significant. Table I and Fig. 1 give the sweep: dispersion is an inverted U peaking at $\beta = 0.5$ and falling to 0.0248 at $\beta = 1$, a 71% reduction, while accuracy rises monotonically in β . H2a and H2b are both Holm-significant wins (Table II). The $\beta = 1$ F1 of 0.891 reproduces the canonical value of the previous paper [6], an internal consistency check obtained without additional cost.

B. Faithfulness fails, decisively

Table II reports the confirmatory family and Fig. 2 the per-seed agreements. Three results combine.

On Kaggle CCF, **H3 is significantly reversed**: the per-sample Spearman correlation between the entropy ranking and the model-aware occlusion ranking is -0.299 in median (per-seed range -0.390 to -0.146), Holm-significant. Entropy does not merely fail to agree with what drives decisions; it points systematically the other way.

On the synthetic track the comparison is cleaner still, because the informative features are known and the quantum

TABLE II
CONFIRMATORY FAMILY: TWO-SIDED WILCOXON SIGNED-RANK TESTS ON PAIRED PER-SEED STATISTICS, ONE HOLM-BONFERRONI CORRECTION, FAMILYWISE $\alpha = 0.05$. A WIN REQUIRES SIGNIFICANCE AND THE PREDICTED DIRECTION.

Comparison	Median Δ	Raw p	Verdict
H2a Dispersion β^* vs 1	+0.0613	0.0020	win
H2b Dispersion 0.25 vs 1	+0.0555	0.0020	win
H3 Entropy vs occlusion	-0.2991	0.0020	reversed
H5a Entropy vs truth	-0.0253	0.0020	reversed
H5b Occlusion vs truth	+0.2368	0.0020	win
H6 F1 β^* vs 1	-0.0137	0.195	no effect

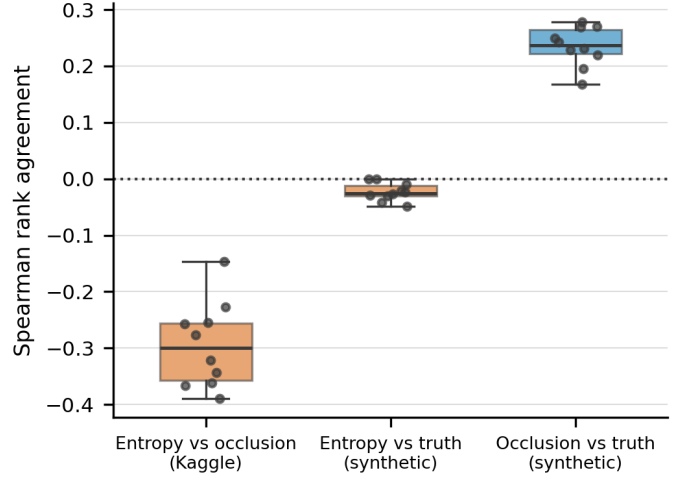


Fig. 2. Per-seed Spearman agreement at $\beta^* = 0.5$. Occlusion recovers synthetic ground truth; entropy does not, and on real data it is anti-correlated with occlusion. Dotted line marks zero agreement.

model demonstrably learns the task: test F1 is 0.921 at β^* against a classical RBF-SVM reference of 0.967. There, **H5b is a win** — occlusion recovers ground truth with median Spearman $+0.237$ — while **H5a is significantly reversed**, entropy attaining -0.025 . Same task, same model, same cohorts, same seeds: the failure belongs to the entropy method specifically, not to the attribution framework, not to task difficulty, and not to an unlearned model.

Five of six comparisons are Holm-significant, and every comparison retained all ten effective pairs except H6, which retained eight. The family was therefore fully decidable: no null below arises from an inability of the test to resolve anything.

C. The mechanism: exact model invariance

Fig. 3 shows the model-randomisation check. When the fitted model's labels are shuffled, per-feature entropy changes by a maximum of *exactly zero* in all ten seeds, while occlusion attributions change by 0.272 on average. Entanglement entropy cannot distinguish a correctly trained model from one fitted to randomised labels, so it fails the sanity check of [11] by construction. This was predeclared, and it is the mechanism behind the two reversals above: a quantity that does not depend on the model cannot track the model's decisions, and any correlation it shows with them is incidental to the encoding's geometry.

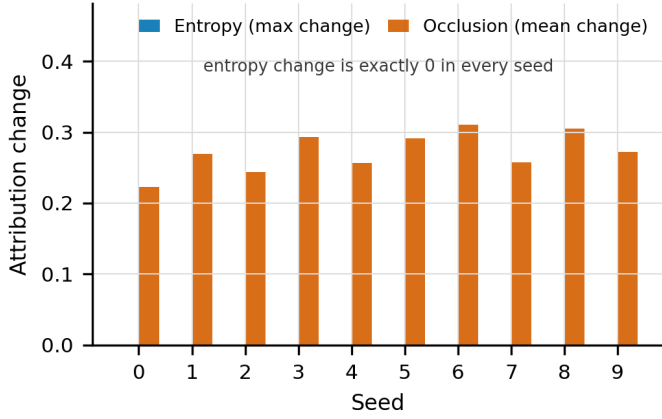


Fig. 3. Change in each attribution when the fitted model’s labels are randomised. Occlusion responds; entropy does not move at all, in any seed.

D. Cost and the accuracy of the auditable regime

H6 finds no detectable accuracy cost to operating at β^* rather than $\beta = 1$ (median $\Delta F1 = -0.0137$, $p = 0.195$), while dispersion there is $3.4\times$ higher. A less degenerate encoding is therefore available essentially free, though Sections V-B and V-C show that what it yields is still not a faithful explanation.

Encoding-level attribution costs one state preparation for all seven features, since every reduced state is obtained from the same statevector. Occlusion costs one kernel row per feature against the fitting cohort — 7×320 entries per explained transaction here — so the cheap quantity is roughly three orders of magnitude cheaper than the faithful one. That asymmetry is precisely what would have made entropy attractive had it worked.

VI. DISCUSSION

A. What was established

Three findings survive correction. Entropy dispersion collapses by 71% at the accuracy-selected operating point, and significantly so. Entanglement entropy is not a faithful attribution: it is anti-correlated with model-aware attribution on real data and significantly fails to recover known ground truth on a synthetic task where a model-aware baseline succeeds. And a more dispersed encoding is available at no detectable accuracy cost.

The three fit together through the model-randomisation result. Because S_j depends only on the feature map and the input, it is constant under any change to the trained model — exactly zero change, in every seed. A quantity with that property can describe how an encoding distributes information, but it cannot, even in principle, explain a particular model’s decisions. The two reversals are then unsurprising in sign and informative in magnitude: what the encoding entangles most is, on this data, close to the opposite of what the classifier relies on.

This is worth stating plainly because the appeal of the idea is real. The cost asymmetry is roughly three orders of magnitude, the quantity is intrinsic rather than imported, and it requires no measurement. Those advantages are all preserved in the

negative result — they simply attach to a diagnostic of the encoding rather than to an explanation of the model.

B. Relation to the standing objection

Rudin’s argument against post-hoc explanation in high-stakes settings [14] is usually answered by claiming a particular method is faithful enough. Our result strengthens the objection in the quantum case rather than rebutting it: the most natural intrinsic candidate fails the most elementary sanity check by construction, and the faithful alternative we tested is the expensive, model-aware one. Under Articles 13 and 86 of the EU AI Act, an explanation offered for a high-risk credit or fraud decision must describe that decision [7], [8]. An encoding diagnostic does not, however elegant its derivation. We make no compliance claim about any system; we observe only that this particular quantity could not support one.

C. Why the amendment mattered

The initial run’s synthetic track underfitted, and its two ground-truth comparisons were inconclusive rather than negative. Had we reported them as nulls, the paper’s central claim — that the failure is specific to entropy — would have rested on the Kaggle agreement result alone, which cannot distinguish “entropy is unfaithful” from “occlusion is also unfaithful and the two disagree”. Amending the cohort construction, with the amendment dated before the run and both runs reported, converted an uninformative null into a decisive contrast. The real-data results are bit-identical across the two runs, so the amendment cannot have moved the findings the main claims rest on.

D. Limitations

Seven bound these claims. First, ten seeds are repeated splits of one dataset, so intervals quantify protocol stability rather than inference over populations. Second, all cohorts are balanced by design, so every metric is a case-control benchmark value. Third, H1’s refutation depends on $\varepsilon = 10^{-3}$; a looser threshold would classify more features as saturated, which is why it was fixed in advance and not revisited. Fourth, occlusion is one model-aware attribution among several, and a different target might agree with entropy better — though nothing could rescue a quantity that is exactly model-invariant. Fifth, the synthetic generator is a single construction, and ground-truth recovery on it does not establish behaviour on all tasks. Sixth, all evaluation is exact statevector simulation; the finite-shot behaviour of either attribution was not measured, and shot noise would affect the expensive one more. Seventh, we have tested one entanglement-based quantity on one feature-map family: mutual information, multi-qubit groupings, or a different encoding could behave differently, though the model-invariance argument applies to any purely encoding-level quantity.

E. What would be worth trying next

Two directions follow directly. A hybrid attribution could use entropy as a cheap prefilter and spend occlusion’s kernel

rows only where the encoding suggests structure — but our results give no reason to expect the prefilter to be informative, and testing that is the point. More promising, the gate-level Shapley approach [1] attributes to circuit components rather than inputs and is not model-invariant, making it a better-founded candidate for quantum-native explanation than the quantity tested here.

VII. CONCLUSION

We asked whether entanglement entropy — intrinsic to a quantum encoding, free of measurement, and obtainable for every feature from a single state preparation — is a faithful feature attribution for a financial-transaction feature map, and we answered under a protocol predeclared down to its thresholds, its comparison family, its claim rule, and its posture on failure.

It is not. On real data the entropy ranking is significantly *anti*-correlated with model-aware occlusion attribution (median Spearman -0.299). On a synthetic task whose informative features are known by construction, and which the quantum model demonstrably learns, occlusion recovers ground truth ($+0.237$) while entropy does not (-0.025), both significant after Holm–Bonferroni correction. The cause is exact: under randomisation of the fitted model’s labels, per-feature entropy changes by precisely zero in all ten seeds while occlusion changes substantially, so entanglement entropy cannot distinguish a trained model from a randomised one and fails the standard sanity check by construction.

Separately, the operating point that accuracy selection chooses compresses even this non-faithful signal: entropy dispersion is 71% lower at $\beta = 1$ than at $\beta = 0.5$, significantly. A regime with $3.4\times$ the dispersion is available at no statistically detectable accuracy cost, which is a useful practical fact about the encoding — but it does not make the resulting entropies explanations.

We report this refutation as predeclared. The contribution we would emphasise is not that one candidate method failed, but that the failure is *localised and mechanistic*: the paper identifies model invariance as the property that disqualifies any purely encoding-level quantity from serving as a model explanation, and demonstrates on a controlled task that a model-aware alternative succeeds where the intrinsic one does not. Quantum models will need explanations that depend on the model, and the cheapness of quantities that do not is not a reason to accept them.

REFERENCES

- [1] R. Heese, T. Gerlach, S. Mücke, S. Müller, M. Jakobs, and N. Piatkowski, “Explaining quantum circuits with Shapley values: Towards explainable quantum machine learning,” *Quantum Machine Intelligence*, vol. 7, no. 1, p. 27, 2025, <https://doi.org/10.1007/s42484-025-00254-8>. [Online]. Available: <https://doi.org/10.1007/s42484-025-00254-8>
- [2] L. Pira and C. Ferrie, “On the interpretability of quantum neural networks,” *Quantum Machine Intelligence*, vol. 6, no. 2, p. 52, 2024, <https://doi.org/10.1007/s42484-024-00191-y>. [Online]. Available: <https://doi.org/10.1007/s42484-024-00191-y>
- [3] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017, advances in Neural Information Processing Systems 30. <https://doi.org/10.48550/arXiv.1705.07874>. [Online]. Available: <https://doi.org/10.48550/arXiv.1705.07874>
- [4] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why Should I Trust You?”: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, <https://doi.org/10.1145/2939672.2939778>. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [5] J. M. V. Sindayigaya, “FT-FM: A financial-transaction-specific quantum feature map for variational quantum classifiers in fraud detection,” *International Journal of Electronics and Telecommunications*, vol. 72, no. 3, pp. 1–8, 2026, <https://doi.org/10.24425/ijet.2026.1729>. [Online]. Available: <https://doi.org/10.24425/ijet.2026.1729>
- [6] —, “FT-FM-lite: Tunable-entanglement refinement and empirical benchmarking on financial-cybersecurity datasets,” 2026, manuscript submitted for publication; no DOI assigned.
- [7] European Parliament and Council, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act),” 2024, oJ L, 2024/1689, 12.7.2024; CELEX 32024R1689. Regulations are not DOI-registered; no DOI assigned. <http://data.europa.eu/eli/reg/2024/1689/oj>.
- [8] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, “Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making,” *IEEE Computational Intelligence Magazine*, vol. 17, no. 1, pp. 72–85, 2022, <https://doi.org/10.1109/MCI.2021.3129960>. [Online]. Available: <https://doi.org/10.1109/MCI.2021.3129960>
- [9] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bannetot, S. Tabak, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020, <https://doi.org/10.1016/j.inffus.2019.12.012>. [Online]. Available: <https://doi.org/10.1016/j.inffus.2019.12.012>
- [10] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017, <https://doi.org/10.48550/arXiv.1702.08608>. [Online]. Available: <https://doi.org/10.48550/arXiv.1702.08608>
- [11] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” 2018, advances in Neural Information Processing Systems 31; no publisher DOI issued. <https://doi.org/10.48550/arXiv.1810.03292>. [Online]. Available: <https://doi.org/10.48550/arXiv.1810.03292>
- [12] D. Alvarez-Melis and T. S. Jaakkola, “On the robustness of interpretability methods,” 2018, <https://doi.org/10.48550/arXiv.1806.08049>. [Online]. Available: <https://doi.org/10.48550/arXiv.1806.08049>
- [13] G. Hooker, L. Mentch, and S. Zhou, “Unrestricted permutation forces extrapolation: Variable importance requires at least one more model,” *Statistics and Computing*, vol. 31, no. 6, p. 82, 2021, <https://doi.org/10.1007/s11222-021-10057-z>. [Online]. Available: <https://doi.org/10.1007/s11222-021-10057-z>
- [14] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, <https://doi.org/10.1038/s42256-019-0048-x>. [Online]. Available: <https://doi.org/10.1038/s42256-019-0048-x>
- [15] F. Mercaldo, G. Ciaramella, G. Iadarola, M. Storto, F. Martinelli, and A. Santone, “Towards explainable quantum machine learning for mobile malware detection and classification,” *Applied Sciences*, vol. 12, no. 23, p. 12025, 2022, <https://doi.org/10.3390/app122312025>. [Online]. Available: <https://doi.org/10.3390/app122312025>
- [16] C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher, “Concentrating partial entanglement by local operations,” *Physical Review A*, vol. 53, no. 4, pp. 2046–2052, 1996, <https://doi.org/10.1103/PhysRevA.53.2046>. [Online]. Available: <https://doi.org/10.1103/PhysRevA.53.2046>
- [17] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010, <https://doi.org/10.1017/CBO9780511976667>. [Online]. Available: <https://doi.org/10.1017/CBO9780511976667>
- [18] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995, <https://doi.org/10.1007/BF00994018>. [Online]. Available: <https://doi.org/10.1007/BF00994018>
- [19] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, “Supervised learning with quantum-enhanced feature spaces,” *Nature*, vol. 567, pp. 209–212, 2019, <https://doi.org/10.1038/s41586-019-0980-2>. [Online]. Available: <https://doi.org/10.1038/s41586-019-0980-2>
- [20] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with undersampling for unbalanced classification,” in *Proc. IEEE Symposium Series on Computational Intelligence*, 2015, pp. 159–166, <https://doi.org/10.1109/SSCI.2015.33>. [Online]. Available: <https://doi.org/10.1109/SSCI.2015.33>

- [21] S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, “Exponential concentration in quantum kernel methods,” *Nature Communications*, vol. 15, p. 5200, 2024, <https://doi.org/10.1038/s41467-024-49287-w>. [Online]. Available: <https://doi.org/10.1038/s41467-024-49287-w>
- [22] J. M. V. Sindayigaya, “Q-RAD: A gated residual architecture for quantum-assisted fraud detection,” 2026, manuscript in preparation; no DOI assigned.
- [23] Qiskit contributors, “Qiskit: An open-source framework for quantum computing,” 2023, <https://doi.org/10.5281/zenodo.2573505>. [Online]. Available: <https://doi.org/10.5281/zenodo.2573505>
- [24] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945, <https://doi.org/10.2307/3001968>. [Online]. Available: <https://doi.org/10.2307/3001968>
- [25] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979, <https://doi.org/10.2307/4615733>. [Online]. Available: <https://doi.org/10.2307/4615733>