

TN-SHAP-Q: Exact Multilinear Extensions of Quantum Neural Networks for Shapley Attribution of Features and Gates

Farzaneh Heidari^{1,2}, Daniel Bilsborrow^{1,2}, Lirandë Pira³, Guillaume Rabusseau^{1,2,4}

¹Mila – Quebec AI Institute ²Université de Montréal

³National University of Singapore ⁴CIFAR AI Chair

Abstract

We introduce TN-SHAP-Q, a multilinear framework for computing Shapley values and interactions in quantum neural networks (QNNs). For QNNs with *finite-lift* data encodings, we show that the circuit output is multilinear in the lifted coordinates and, moreover, coincides exactly with the *multilinear extension* (MLE) of the induced cooperative game. This connection unifies feature and gate attribution within a common game-theoretic framework, and allows attribution to be computed directly from the multilinear structure of the model, without explicit coalition enumeration or surrogate fitting. Three consequences follow from this exact MLE realization. First, first-order Shapley values are computed exactly by Owen diagonal integration. Second, order- k interaction indices are obtained from the same extension, with a Gauss–Legendre exactness threshold $M \geq \lceil (P - k + 1)/2 \rceil$ that decreases with interaction order, where P is the number of players (features or gates). Third, exact first-order Shapley values are recovered with polynomially many extension evaluations $O(P^2)$ rather than 2^P evaluations of coalition enumeration. Experiments on a trained Iris QNN, a breast-cancer variational quantum classifier, synthetic QNNs and the `qshaptools` QAOA benchmark reproduce exact Shapley values and interactions to machine precision, validate the predicted quadrature-exactness thresholds, exhibit the expected $1/\sqrt{N}$ degradation under finite-shot noise (where N denotes the measurement shots) and reveal interaction structure that first-order attribution misses.

1 Introduction

QNN predictions remain hard to attribute to input features or circuit components. Existing quantum explainability methods fall into two families. Local surrogate explainers such as Q-LIME [12] fit a linear model near an instance and use its local coefficients as explanations, resulting in a different notion of importance from game-theoretic Shapley attributions. Shapley-based methods [7] instead treat features or gates as players in a cooperative game [14] and recover the value either by coalition enumeration, which scales as 2^P in the number of players P , or by Monte Carlo sampling, which reduces the computational cost while introducing approximation error. We observe that finite-lift QNNs, QNNs with finite-dimensional encodings [3, 13], already realize the multilinear extension (MLE) of the induced cooperative game. Consequently, Shapley values and interaction indices can be recovered directly through Owen’s multilinear-extension formulation [11]. Using Gauss–Legendre quadrature [1], exact first-order Shapley values are obtained with polynomially many extension evaluations rather than exponential coalition enumeration, a classical reduction in circuit evaluations relative to enumeration, not a quantum runtime speedup. TN-SHAP-Q extends the TN-SHAP family [8, 9] to quantum models, where finite-lift QNNs provide direct access to the exact MLE. **Contributions.** (i) We prove that feature and gate value functions in finite-lift QNNs realize the exact multilinear extension of the induced game. (ii) We derive exact first-order Shapley values with polynomial query complexity and exact interaction indices from the same multilinear extension. (iii) We show extension evaluations are unbiased under finite-shot estimation. (iv) We validate the theory on classification and QAOA benchmarks, recovering exact first-order Shapley attributions and interactions to machine precision.

2 Preliminaries

A QNN encodes feature x_i by a single $R_Y(x_i)$ rotation, applies an input-independent block U , measures a Hermitian observable O , returning $f(x) = \text{Tr} \left(O U \left(\bigotimes_{i=1}^d \rho_i(x_i) \right) U^\dagger \right)$. For a single R_Y encoding, $\rho_i(x_i) = \frac{1}{2}(I + \cos x_i Z + \sin x_i X)$, so the encoding depends on x_i only through the finite-dimensional feature vector (lifted coordinates) $\phi_i(x_i) = [1, \cos x_i, \sin x_i]$. More generally, we require only a *product encoding* with a finite per-feature expansion $\rho_i(x_i) = \sum_{k=1}^{r_i} \phi_{ik}(x_i) A_{ik}$, where A_{ik} are fixed, input-independent basis operators and $\phi_{ik}(x_i)$ are the corresponding lifted coordinates. All trainable and entangling layers are contained in the

input-independent block U ; they modify the coefficient tensor but not the multilinear form. A more detailed definition of the circuit is given in App. A. The construction applies to QNNs with a single product-encoding layer. Data re-uploading architectures, where a feature is re-encoded across layers interleaved with trainable gates, fall outside this setting: the masking interpolation is no longer multilinear in the coalition variable (App. F). We study two cooperative games on such circuits. In the **feature game**, players are input features. Given a baseline b , the coalition value is $v_x(S) = f(x_S)$, where features in S are kept and features outside S are replaced by baseline values. In the **gate game**, players are gates of $U = g_m \cdots g_1$. For a coalition S of retained gates, $v_g(S) = \text{Tr}(O U_S \rho U_S^\dagger)$, which matches the SVQX/qshaptools gate-attribution setting. A more detailed definition of Shapley values and interaction indices is given in App. B.

3 Multilinear Extension

TN-SHAP-Q treats feature and gate attribution as cooperative games (Sec. 2) and computes their Shapley values through each game’s *multilinear extension* (Def. 1, Thm. 1, Props. 1–3).

Definition 1 (Multilinear extension and Owen value). *Let $\mathcal{P} = \{1, \dots, P\}$ denote the player set, where $P = |\mathcal{P}|$. The multilinear extension (MLE) of a game $v : 2^{\mathcal{P}} \rightarrow \mathbb{R}$ on players $\mathcal{P}$ is $F(z) = \sum_{S \subseteq \mathcal{P}} v(S) \prod_{i \in S} z_i \prod_{i \notin S} (1 - z_i)$, $z \in [0, 1]^P$; it is the unique multilinear function agreeing with v on every coalition vertex, $F(\mathbf{1}_S) = v(S)$. The first-order Shapley value is the diagonal Owen integral $\Phi_i = \int_0^1 \partial_i F(t\mathbf{1}) dt$, and order- k interaction indices are the diagonal integrals of mixed partials, $I(S) = \int_0^1 \partial_S F(t\mathbf{1}) dt$, $\partial_S = \prod_{i \in S} \partial_i$ [6, 5].*

Theorem 1 (Exact realization of multilinear extension). (a) Features. *With lifted instance $u_i = \phi_i(x_i)$ and baseline $w_i = \phi_i(b_i)$, the function $F_x(z) = f(w_1 + z_1(u_1 - w_1), \dots, w_d + z_d(u_d - w_d))$ is multilinear in $z \in [0, 1]^d$ and satisfies $F_x(\mathbf{1}_S) = v_x(S)$.* (b) Gates. *Letting gate i act with weight z_i via the channel $\mathcal{E}_i^{z_i}(\rho) = z_i g_i \rho g_i^\dagger + (1 - z_i)\rho$, the value $F_g(z)$ is multilinear in z and satisfies $F_g(\mathbf{1}_S) = v_g(S)$.*

By uniqueness of the MLE, F_x/F_g is the multilinear extension of v_x/v_g , realized by the circuit. Arbitrary trainable and entangling layers change only the coefficient tensor.

Proposition 1 (First-order exactness and cost). *$\partial_i F(t\mathbf{1}) = F(z_i=1, z_{-i}=t) - F(z_i=0, z_{-i}=t)$ is a polynomial in t of degree $\leq P-1$. Hence M -point Gauss–Legendre quadrature computes Φ_i exactly whenever $M \geq \lceil P/2 \rceil$, using $2PM$ evaluations of F versus 2^P for enumeration.*

Proposition 2 (Interaction exactness). *For multilinear F on P variables and a coalition S with $|S| = k$, the diagonal mixed partial $\partial_S F(t\mathbf{1})$ is a polynomial in t of degree $\leq P-k$. Hence M -point Gauss–Legendre quadrature computes the order- k interaction index $I(S) = \int_0^1 \partial_S F(t\mathbf{1}) dt$ exactly whenever $M \geq \lceil (P-k+1)/2 \rceil$, using $2^k M$ extension evaluations per coalition. The required quadrature order therefore decreases with interaction order.*

Proposition 3 (Unbiased finite-shot estimation). *Let $\hat{F}_N(z)$ be the N -shot estimate of an extension value $F(z)$, obtained by classical randomization over coalition vertices (sample $z' \in \{0, 1\}^P$ with $z'_i \sim \text{Bernoulli}(z_i)$ independently, prepare the corresponding circuit, and average the observable over N shots). Then $\mathbb{E}[\hat{F}_N(z)] = F(z)$ for every $z \in [0, 1]^P$. Consequently the Owen estimator $\hat{\Phi}_i = \sum_n w_n \hat{F}_N(z^{(n)})$, a fixed linear combination of extension evaluations, is unbiased, $\mathbb{E}[\hat{\Phi}_i] = \Phi_i$, with error decaying as $O(1/\sqrt{N})$.*

Proofs of Thm. 1 and Props. 1–2 are in App. C; the finite-lift circuit and the cooperative-game definitions are in Apps. A and B. Proof of Prop. 3 is in App. C.

Each evaluation of F is therefore preparable on hardware by classical randomization (sample a coalition vertex $z \in \{0, 1\}^P$ according to the corresponding probabilities), so the method is not restricted to state vector simulation; finite-shot estimation makes $\hat{F}$ an unbiased estimate of F and the quadrature a fixed linear combination of such estimates (Sec. 4). We defer the Limitations section to App. G.

4 Experiments

We use exact state vector simulation for ground truth (2^P enumeration) and add a finite-shot study; further details of the experiment settings are in App. D.

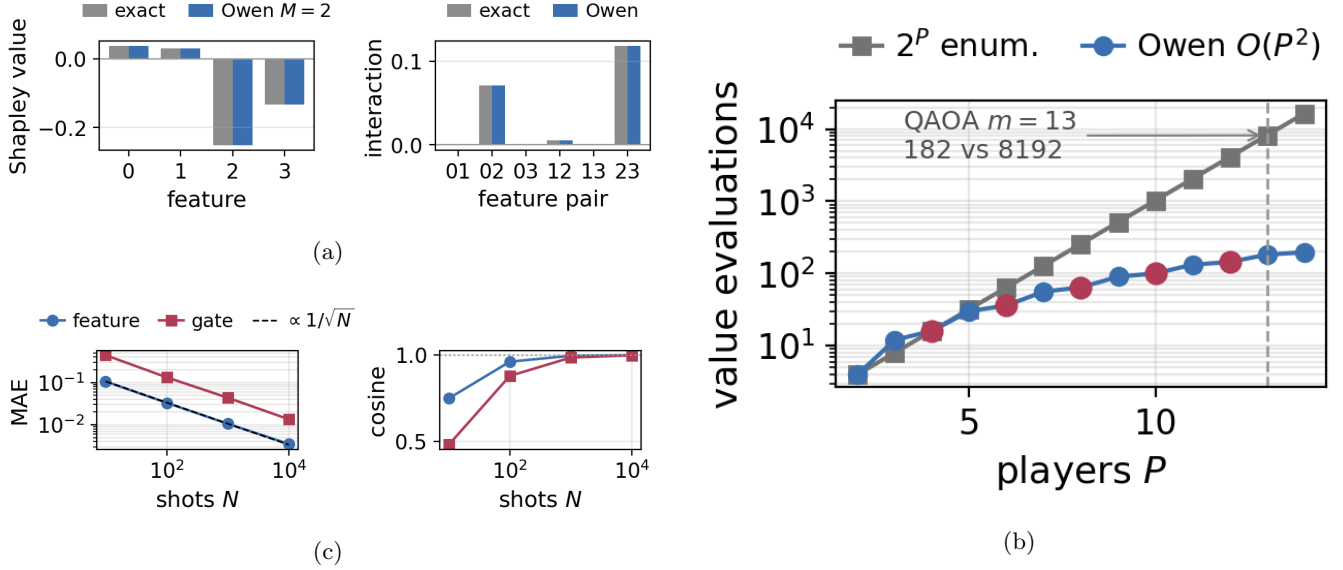


Figure 1: (a) First-order and order-2 feature attribution on the Iris QNN ($d=4$): Owen ($M=2$) coincides with enumeration; the dominant pairwise coupling is a redundancy additive values miss. (b) Exact $O(P^2)$ vs $O(2^P)$ evaluation count (combinatorial), validated by exact recovery on synthetic QNNs for $d=4, \dots, 12$; the dashed line marks the $m=13$ QAOA example (182 vs 8192). (c) Finite-shot robustness: MAE and cosine vs shot budget $N \in \{10, 10^2, 10^3, 10^4\}$, error $\propto 1/\sqrt{N}$.

4.1 Feature attribution and the exactness threshold

A single- R_Y QNN on Iris (setosa vs. versicolor, $d=4$, train acc. 0.97) explained against a dataset-mean baseline (b_i in Prop. 1) reproduces exact feature Shapley values to MAE 4.2×10^{-17} (cosine 1.000) at $2dM=16$ evaluations with $M=2$ coinciding with 2^d enumeration (Fig. 1a, left); $M=1$ is inexact (MAE 1.8×10^{-4}), confirming the $M \geq \lceil d/2 \rceil$ threshold of Prop. 1. The same extension yields the 6 exact pairwise interaction indices (MAE 9.5×10^{-17}); the strongest, features 2–3 ($I_{ij} = +0.119$), is a redundancy missed by additive attribution (Fig. 1a, right). The same threshold holds on a stronger, non-toy benchmark—a single- R_Y VQC on Wisconsin breast cancer (top-6 features, train acc. 0.92; Table 1, row 08), where $M=3$ reproduces the exact 6 Shapley values (MAE 5.7×10^{-17} , 36 vs 64 evaluations) while $M=2$ is inexact.

4.2 Dimension scaling

On synthetic single- R_Y QNNs with $d = 4, 6, 8, 10, 12$, TN-SHAP-Q recovers the exact Shapley values at every d (max error 10^{-16}) while its $2d\lceil d/2 \rceil$ cost separates sharply from 2^d (144 vs 4096 at $d=12$; Fig. 1b), validating Prop. 1 and the query-complexity bound of Table 2.

4.3 Shot-noise robustness

Replacing exact expectations by N -shot estimates of each extension value, the Owen estimator stays unbiased and its MAE, cosine, and rank correlation against the exact values improve as $1/\sqrt{N}$ across $N \in \{10, 10^2, 10^3, 10^4\}$ for feature and gate games (fitted log-log MAE -0.49 and -0.50); Fig. 1c).

4.4 Gate attribution and interactions

On the `qshaptools` example [7] (Hamiltonian $H = Z_1 Z_2 + 2Z_2 - 3Z_0 Z_2$, a QAOA ansatz decomposed to $m=13$ gate players; Table 1, rows 05–06), TN-SHAP-Q reproduces the exact gate Shapley to max error 4.1×10^{-15} using $2m\lceil m/2 \rceil=182$ extension evaluations versus $2^{13}=8192$ SVQX value evaluations ($\approx 45\times$ fewer), with the gap growing as $O(m^2)$ vs $O(2^m)$ (marked at $m=13$ in Fig. 1b; Table 2). The channel-mixture extension also gives all $\binom{13}{2}=78$ pairwise gate interactions to max error 4.4×10^{-15} (24 extension evaluations per pair at the $M=6$ guaranteed by Prop. 2); single- and double-layer ablations leave $\langle H \rangle \approx 0$ ($|\langle H \rangle| \leq 6 \times 10^{-16}$ vs 1.81 for the full circuit), so the energy is realized only by the full init $\times$ cost $\times$ mixer combination—interaction-dominated (pairwise interaction magnitude $4.6\times$ the first-order), and it is not additive; the interaction heatmap and per-gate breakdown are in Fig. 4 (App. E.2).

References

- [1] P.J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Dover Books on Mathematics Series. Dover Publications, 2007.
- [2] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv: Quantum Physics*, 2014.
- [3] Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors, 2018. arXiv:1802.06002.
- [4] Rory A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7:179–188, 1936.
- [5] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.
- [6] John C. Harsanyi. A simplified bargaining model for the n -person cooperative game. *International Economic Review*, 4(2):194–220, 1963.
- [7] Raoul Heese, Thore Gerlach, Sascha Mücke, Sabine Müller, Matthias Jakobs, and Nico Piatkowski. Explaining quantum circuits with shapley values: Towards explainable quantum machine learning. *Quantum Machine Intelligence*, 7:27, 2025.
- [8] Farzaneh Heidari, Chao Li, and Guillaume Rabusseau. Tractable shapley values and interactions via tensor networks. In *Proc. 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, 2026. arXiv:2510.22138.
- [9] Farzaneh Heidari and Guillaume Rabusseau. TN-SHAP-G: Graph-structured tensor network surrogates for shapley values and interactions. In *Proc. 43rd International Conference on Machine Learning (ICML)*, PMLR, 2026. arXiv:2606.01540.
- [10] Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, Blake R. Johnson, and Jay M. Gambetta. Quantum computing with qiskit. *ArXiv*, abs/2405.08810, 2024.
- [11] Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):P64–P79, 1972.
- [12] Lirandë Pira and Chris Ferrie. On the interpretability of quantum neural networks. *Quantum Machine Intelligence*, 6:52, 2024.
- [13] 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. arXiv:2008.08605.
- [14] Lloyd S. Shapley. A value for n -person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games II*, number 28 in Annals of Mathematics Studies, pages 307–317. Princeton University Press, 1953.
- [15] William H. Wolberg, Olvi L. Mangasarian, and W. Nick Street. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993.

A Finite-lift QNN

The QNN encodes feature x_i through a single $R_Y(x_i)$ rotation, applies an input-independent variational block $U(\theta)$ (arbitrarily deep, trainable and entangling), and measures an observable O ,

$$f(x) = \text{Tr}\left(O U \left(\bigotimes_{i=1}^d \rho_i(x_i)\right) U^\dagger\right), \quad \rho_i(x_i) = \frac{1}{2}(I + \cos x_i Z + \sin x_i X),$$

so each feature carries the finite lift $\phi_i(x_i) = [1, \cos x_i, \sin x_i]$ ($r_i=3$). All input dependence enters through the single-qubit encoding layer; U and O are input-independent (Fig. 2). More generally we require only a product encoding with a finite per-feature expansion $\rho_i(x_i) = \sum_{k=1}^{r_i} \phi_{ik}(x_i) A_{ik}$.

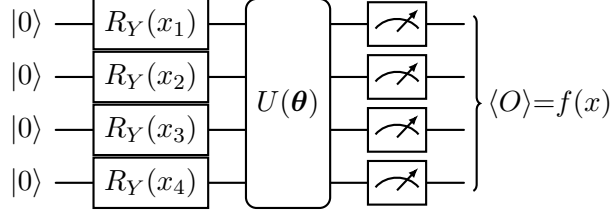


Figure 2: Finite-lift QNN ($d=4$). Each feature is encoded by one $R_Y(x_i)$ into $\rho_i(x_i) = \frac{1}{2}(I + \cos x_i Z + \sin x_i X)$, the lift $\phi_i(x_i) = [1, \cos x_i, \sin x_i]$. The product state passes through an input-independent $U(\theta)$ and O is measured; since all input dependence is in the encoding layer, f is multilinear in the lifted coordinates (App. C).

B Cooperative-game definitions

Definition 2 (Shapley value). *Let $\mathcal{P} = \{1, \dots, P\}$ denote the player set. For a game $v : 2^{\mathcal{P}} \rightarrow \mathbb{R}$ the Shapley value of player i is $\Phi_i(v) = \sum_{S \subseteq \mathcal{P} \setminus \{i\}} \frac{|S|!(P-|S|-1)!}{P!} (v(S \cup \{i\}) - v(S))$.*

Definition 3 (Owen diagonal integral). *Let F be the multilinear extension of v , $\Phi_i = \int_0^1 \partial_i F(t\mathbf{1}) dt$.*

Definition 4 (Shapley interaction index). *The Grabisch–Roubens order- $|S|$ interaction index [5] is $I(S) = \int_0^1 \partial_S F(t\mathbf{1}) dt$, $\partial_S = \prod_{i \in S} \partial_i$, which reduces to the Shapley value when $|S|=1$.*

C Proofs

These establish Thm. 1 and Props. 1–3 of the main text.

(a) Features — Thm. 1(a). [Feature multilinearity] Under the finite product encoding of App. A,

$$f(x) = \sum_{a_1, \dots, a_d} T[a_1, \dots, a_d] \prod_{i=1}^d \phi_{i,a_i}(x_i),$$

with the input-independent tensor $T[a_1, \dots, a_d] = \text{Tr}(O U (\bigotimes_i A_{i,a_i}) U^\dagger)$; hence f is multilinear in the lift.

Proof. Substitute the expansions into f and expand by distributivity,

$$\bigotimes_i \left(\sum_{a_i} \phi_{i,a_i}(x_i) A_{i,a_i} \right) = \sum_{a_1, \dots, a_d} \left(\prod_i \phi_{i,a_i}(x_i) \right) \bigotimes_i A_{i,a_i}.$$

Linearity of $\text{Tr}(O U \cdot U^\dagger)$ pulls the scalars out, giving the stated tensor form; each ϕ_{i,a_i} depends on a single x_i , so f is multilinear (degree ≤ 1 per feature group). Changing U changes only the coefficient tensor T , leaving the multilinear form unchanged. $\square$

Corollary 1 (Feature-game MLE — realizes Thm. 1(a)). *Fix instance x , baseline b , lifts $u_i = \phi_i(x_i)$, $w_i = \phi_i(b_i)$, and the affine interpolation $\psi_i(z_i) = w_i + z_i(u_i - w_i)$. Here $\psi_i(z_i)$ denotes the interpolated lifted coordinates, not a quantum state. Then $F_x(z) = f(\psi_1(z_1), \dots, \psi_d(z_d))$ is multilinear in $z \in [0, 1]^d$ and equals the multilinear extension of v_x , with $F_x(\mathbf{1}_S) = v_x(S)$.*

Proof. f is multilinear in the lift (shown above) and each ψ_i is affine in z_i ; composing a multilinear map with per-coordinate affine maps preserves multilinearity, so F_x is multilinear. At a vertex $\mathbf{1}_S$, $\psi_i = u_i$ for $i \in S$ and $\psi_i = w_i$ otherwise, hence $F_x(\mathbf{1}_S) = f(x_S) = v_x(S)$. The MLE is the unique multilinear function matching v_x on all vertices, so F_x is that extension. $\square$

(b) Gates — Thm. 1(b). [Gate multilinearity] Let $U = g_m \cdots g_1$ and let gate i act with keep-weight z_i through the channel $\mathcal{E}_i^{z_i}(\rho) = z_i g_i \rho g_i^\dagger + (1 - z_i)\rho$. Then $F_g(z) = \text{Tr}(O(\mathcal{E}_m^{z_m} \circ \cdots \circ \mathcal{E}_1^{z_1})(\rho))$ is multilinear in $z \in [0, 1]^m$ and equals the MLE of $v_g(S) = \text{Tr}(O U_S \rho U_S^\dagger)$, with $F_g(\mathbf{1}_S) = v_g(S)$.

Proof. Each $\mathcal{E}_i^{z_i}$ is affine in z_i and independent of $z_{j \neq i}$. Composing and expanding over the binary choices (apply g_i / identity) gives $(\mathcal{E}_m^{z_m} \circ \cdots \circ \mathcal{E}_1^{z_1})(\rho) = \sum_{S \subseteq [m]} (\prod_{i \in S} z_i \prod_{i \notin S} (1 - z_i)) U_S \rho U_S^\dagger$, $U_S = \prod_{i \in S} g_i$. Applying $\text{Tr}(O \cdot)$ and linearity yields $F_g(z) = \sum_S (\prod_{i \in S} z_i \prod_{i \notin S} (1 - z_i)) v_g(S)$; each z_i appears in one factor, so F_g is multilinear, and $F_g(\mathbf{1}_S) = v_g(S)$. This is exactly the MLE of v_g . $\square$

(c) First-order exactness and cost — Prop. 1.

Proof of Prop. 1. F is affine in z_i , so the partial is the exact finite difference $\partial_i F(t\mathbf{1}) = F(z_i=1, z_{-i}=t) - F(z_i=0, z_{-i}=t)$, a polynomial in t of degree $\leq P - 1$ (one factor per other player). An M -point Gauss–Legendre rule is exact for degree $\leq 2M - 1$, so the Owen integral $\int_0^1 \partial_i F(t\mathbf{1}) dt$ is exact once $2M - 1 \geq P - 1$, i.e. $M \geq \lceil P/2 \rceil$. Each node needs F at $z_i \in \{0, 1\}$ (2 evaluations), giving $2PM$ evaluations over P players [14]. $\square$

(d) Interaction exactness — Prop. 2.

Proof of Prop. 2. Write $F(z) = \sum_{T \subseteq \mathcal{P}} c_T \prod_{i \in T} z_i$ (multilinearity makes this exact). Differentiating in the variables of S ($|S|=k$) annihilates every monomial with $T \not\supseteq S$ and strips $\{z_i\}_{i \in S}$ from the rest, so $\partial_S F(z) = \sum_{T \supseteq S} c_T \prod_{i \in T \setminus S} z_i$. Each surviving monomial has $|T \setminus S| \leq P - k$ factors, so $\partial_S F(t\mathbf{1})$ is a polynomial in t of degree $\leq P - k$. An M -point Gauss–Legendre rule is exact for degree $\leq 2M - 1$, hence exact for $I(S) = \int_0^1 \partial_S F(t\mathbf{1}) dt$ once $2M - 1 \geq P - k$, i.e. $M \geq \lceil (P - k + 1)/2 \rceil$. Multilinearity makes $\partial_S F(t\mathbf{1})$ the 2^k -corner finite difference $\sum_{R \subseteq S} (-1)^{k-|R|} F(z_R=1, z_{S \setminus R}=0, z_{\mathcal{P} \setminus S}=t)$, so each of the M nodes costs 2^k evaluations — $2^k M$ per coalition. The threshold therefore *decreases* with interaction order. $\square$

Hardware realizability. For both games the interpolated evaluation point is a convex mixture of physical states: $w_i + z_i(u_i - w_i)$ corresponds to $(1 - z_i)\rho(b_i) + z_i\rho(x_i)$, a valid density matrix, and $\mathcal{E}_i^{z_i}$ is a convex mixture of “apply g_i ” and identity.

(e) Unbiased finite-shot estimation — Prop. 3.

Proof of Prop. 3. Fix $z \in [0, 1]^P$. Conditioned on a sampled vertex $z' \in \{0, 1\}^P$, the N -shot mean of the observable is an unbiased estimate of its expectation at that vertex, $\mathbb{E}[\cdot \mid z'] = F(\mathbf{1}_{S'}) = v(S')$ with $S' = \{i : z'_i=1\}$. The vertices are drawn independently with $\Pr[z'=\mathbf{1}_S] = \prod_{i \in S} z_i \prod_{i \notin S} (1 - z_i)$, so by the tower rule

$$\mathbb{E}[\hat{F}_N(z)] = \mathbb{E}_{z'}[F(\mathbf{1}_{S'})] = \sum_{S \subseteq \mathcal{P}} \left(\prod_{i \in S} z_i \prod_{i \notin S} (1 - z_i) \right) v(S) = F(z),$$

the last equality being the multilinear extension (Def. 1). The Owen estimator is a fixed linear combination $\hat{\Phi}_i = \sum_n w_n \hat{F}_N(z^{(n)})$ of such evaluations, so linearity of expectation gives $\mathbb{E}[\hat{\Phi}_i] = \sum_n w_n F(z^{(n)}) = \Phi_i$. Each $\hat{F}_N$ has variance $O(1/N)$, hence the estimator error scales as $O(1/\sqrt{N})$, matching the empirical slopes -0.49 and -0.50 (Fig. 1c). $\square$

experiments	players	setting	source	we replicate
01,02	features	single- R_Y finite-lift encoding [13]; variational classifier [3]	encoding QNN	finite-lift QNN; Owen/GR attribution [11, 5] vs exact 2^d enumeration
03	features	same encoding [3, 13], synthetic weights		weight-independent exactness, $d=4..12$
04	feat./gates	projective shot noise (standard); motivated by Q-LIME [12]		N -shot estimates, $N \in \{10, \dots, 10^4\}$
05,06	gates	SVQX gate game + QAOA example [7]; QAOA [2]; Qiskit QAOAAnsatz [10]		their gate-removal game and $H=Z_1Z_2+2Z_2-3Z_0Z_2$ example exactly
07	features	Q-LIME local-linear surrogate [12]	surrogate	their surrogate; ranking-only comparison
08	features	single- R_Y VQC [13, 3] on a UCI benchmark [15]		finite-lift VQC on breast cancer; exact recovery vs 2^d

Table 1: Experimental details

D Experimental settings

The single- R_Y finite-lift encoding is the single-frequency case of the data-encoding spectrum of [13] with the variational-classifier form of [3]; the gate game and the $H=Z_1Z_2+2Z_2-3Z_0Z_2$, QAOA example are the SVQX/qshaptools setting of [7] (QAOA [2], Qiskit [10]), reproduced verbatim; Q-LIME is the local-surrogate explainer of [12]. Iris [4] and Wisconsin breast cancer [15] are standard UCI benchmarks. All experiments are reproducible from the public demo, <https://github.com/farzana0/TN-SHAP-Q-demo>.

E Additional experiments

This appendix complements Sec. 4; the appendix follows with comparison with Q-LIME, and the gate-interaction study shows exactness carries to order-2 indices on a structured circuit.

E.1 Q-LIME ranking comparison

Q-LIME [12] fits a kernel-weighted local-linear surrogate to sampled feature masks; its coefficients are not Shapley values, so comparing magnitudes (or an MAE against the exact Shapley value) is ill-posed. We therefore compare only the induced *feature ranking*. On the Iris QNN ($d=4$) its rank agreement with TN-SHAP-Q saturates below 1 (Spearman ≈ 0.95 , Kendall $\tau \approx 0.92$, top-1 ≈ 0.88), see (Fig. 3).

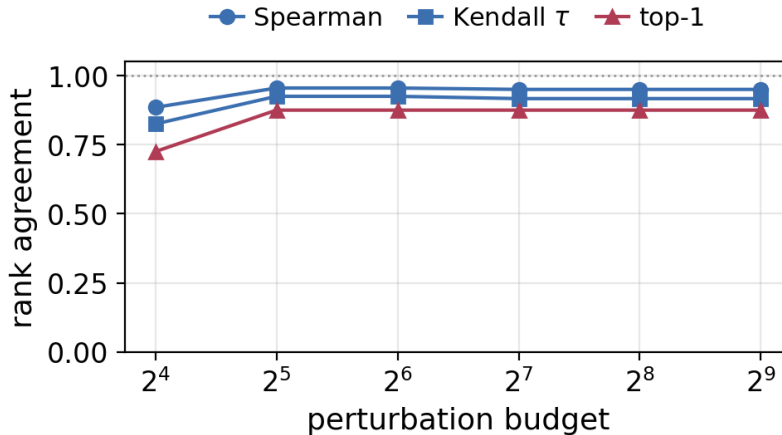


Figure 3: Q-LIME feature ranking vs. TN-SHAP-Q on the Iris QNN ($d=4$): Spearman, Kendall τ , and top-1 agreement as the perturbation budget grows ($16 \rightarrow 512$). Agreement saturates below 1 and does not improve with more perturbations. Magnitudes are not compared (Q-LIME returns local-linear coefficients, not Shapley values).

E.2 Order-2 gate interactions

On the qshaptools QAOA circuit (Sec. 4) the channel-mixture extension gives *exact* pairwise gate interactions (Def. 4): all $\binom{13}{2}=78$ match 2^m Grabisch–Roubens enumeration to max error 4.4×10^{-15} at $2^{|S|}M=24$

Table 2: Query complexity for exact first-order Shapley attribution over P players (M quadrature nodes; k interaction order).

method	exact?	evaluations
coalition enumeration	yes	2^P
TN-SHAP-Q, Owen first-order	yes	$2PM$, $M = \lceil P/2 \rceil$
TN-SHAP-Q, order- k interaction (per coalition)	yes	$2^k M$, $M = \lceil (P - k + 1)/2 \rceil$
SVQX subsampled [7]	no	$\gg$ (MAE $\rightarrow 0$ only in the limit)
Q-LIME [12]	no (ranking-only)	sampling-budget dependent

extension evaluations per pair ($M=6$) — 1872 for all pairs versus $2^{13}=8192$ to form the game. The structure is interaction-dominated (Fig. 4): single- and double-layer ablations leave $\langle H \rangle \approx 0$, so the energy (1.81) is realized only by the full $\text{init} \times \text{cost} \times \text{mixer}$ combination, and the total pairwise interaction magnitude is $4.6 \times$ the first-order.

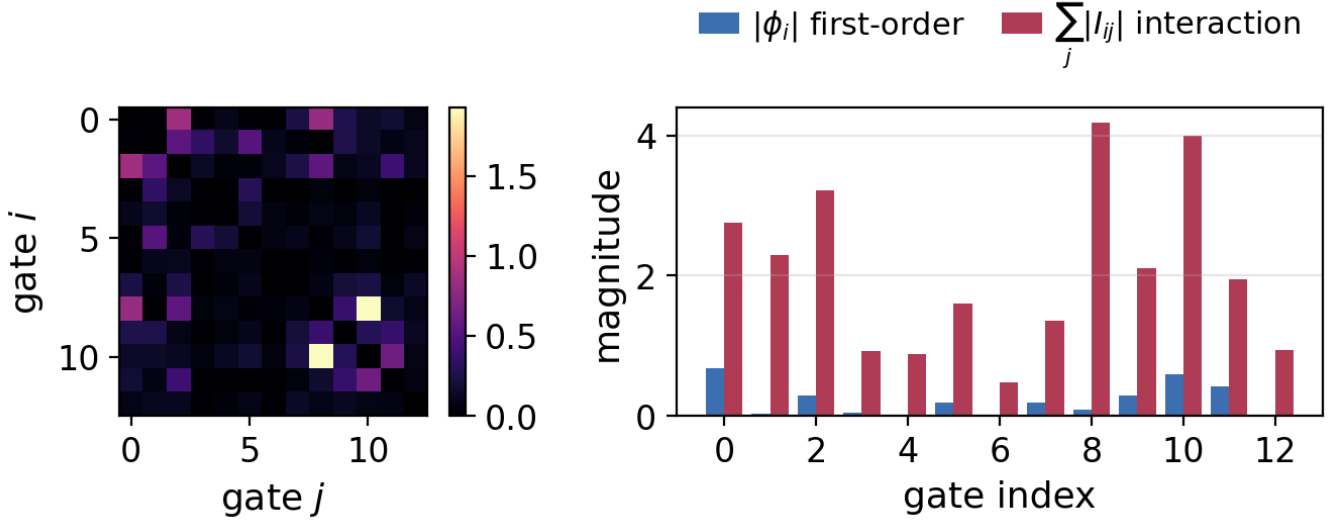


Figure 4: Order-2 gate interactions for the $m=13$ QAOA circuit. (a) $|I(\{i, j\})|$ over gate pairs; (b) per-gate first-order $|\Phi_i|$ vs. total interaction $\sum_j |I_{ij}|$ — a gate with small Φ but large interaction sum is important only in combination. TN-SHAP-Q matches 2^m enumeration to max 4.4×10^{-15} at 24 evaluations/pair.

F Beyond exact finite-lift

When exact extension evaluations are out of reach—deep re-uploading circuits, large gate counts, or black-box quantum simulators returning only sampled values—TN-SHAP-Q reverts to the TN-SHAP procedure [8, 9]: fit a multilinear tensor-network surrogate to sampled extension values and read Shapley values and interactions off the surrogate. Finite-lift QNNs are the setting where this surrogate collapses to the exact circuit response; the tensor-network surrogate extends this to more general settings, we leave the formal proofs and experiments of the general setting to future work.

G Limitations

TN-SHAP-Q evaluates the MLE at interpolated (convex-mixture) inputs, whereas Q-LIME and SVQX evaluate ordinary masked circuits; exactness is established against enumeration. The polynomial-versus-exponential saving is in circuit evaluations relative to coalition enumeration over the same circuit; it is a classical combinatorial saving, not a quantum runtime speedup. The exact-realization result is confined to single-layer product encodings—data re-uploading and non-product encodings leave the regime and require the tensor-network surrogate of Sec. F.