Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Coherent Projector-Overlap QFNNs with Chebyshev Responses: Expressivity, Approximation, and Coherent Depth

Andrej Sum-Shik

asumshik@gmail.com

University of Copenhagen

Research Article

Keywords: Quantum machine learning, quantum neural networks, Chebyshev polynomials, alternating reflections, universal approximation, coherent quantum circuits

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10870861/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Coherent Projector-Overlap QFNNs with Chebyshev Responses: Expressivity, Approximation, and Coherent Depth

Andrej Sum-Shik[1*]

[1]MSc Programme in Quantum Information Science, Faculty of Science, University of Copenhagen, Copenhagen, Denmark.

Corresponding author(s). E-mail(s): asumshik@gmail.com;

Abstract

Variational quantum models combine data encodings, trainable unitaries, and terminal measurements. We introduce a coherent quantum feedforward neural network (QFNN) with an explicit neural-style activation: trainable projector overlaps act as pre-activations, and alternating reflections implement exact Chebyshev responses. For a state $|\psi\rangle$, an orthogonal projector $\boldsymbol{Q}$, and $\boldsymbol{p} = \langle\psi|\,\boldsymbol{Q}\,|\psi\rangle$, $\boldsymbol{d}$ reflection pairs produce $\boldsymbol{T_d(2p-1)}$.

For pure-state inputs, we characterize the complete idealized degree-at-most-$\boldsymbol{D}$ one-layer hypothesis space as $\boldsymbol{\rho \mapsto \mathbf{Tr}(A\rho^{\otimes D})}$ with $\boldsymbol{A \in \mathbf{Herm}(\mathbf{Sym}^D \mathcal{H})}$. This yields its exact dimension, polynomial-fidelity kernel, strict degree hierarchy, domain-restricted capacity, interpolation properties, and finite separation from single-copy observable models. We derive Chebyshev approximation bounds, prove Universal Approximation Theorems for scalar and finite-dimensional continuous targets, and obtain norm- and margin-sensitive generalization bounds.

The construction extends to fully coherent depth without intermediate measurement. The exact activation law and idealized Universal Approximation Theorem persist at every prescribed finite depth, while a process-feature representation and an exact Chebyshev-composition subarchitecture clarify the expressive and resource roles of recursion. Six-seed simulations on eight-class MNIST, Fashion-MNIST, and KMNIST obtain exact test accuracies of $\mathbf{90.57 \pm 0.20\%}$, $\mathbf{74.55 \pm 0.51\%}$, and $\mathbf{67.61 \pm 0.65\%}$, respectively. Together, these results provide an explicit mathematical framework for neural-style polynomial responses in coherent variational quantum architectures.

Keywords: Quantum machine learning, quantum neural networks, Chebyshev polynomials, alternating reflections, universal approximation, coherent quantum circuits

1 Introduction

The classical one-hidden-layer universal-approximation paradigm begins with a simple architecture: finitely many trainable affine pre-activations are passed through a nonlinear response and combined by a linear readout. Under standard assumptions, allowing the hidden width and parameters to vary makes such finite banks dense in broad classes of continuous functions (Cybenko, 1989). This shallow result motivates our first question: can a coherent quantum circuit realize an explicit analogue of a nonlinear hidden unit and a finite one-layer activation bank?

Variational quantum models provide a natural setting for this question: data encodings, trainable circuits, and measurements define architecture-dependent feature maps and hypothesis classes (Schuld and Killoran, 2019; Abbas et al., 2021).

Our aim is to construct a coherent model whose activation law, finite-degree function space, approximation properties, and statistical capacity can be characterized explicitly. We first develop and analyse a one-layer quantum overlap perceptron. The same primitive then admits a recursive implementation in which the complete hidden state is passed coherently from one layer to the next. Motivated by the compositional role of depth in classical neural networks (Mhaskar and Poggio, 2016; Telgarsky, 2016; Eldan and Shamir, 2016; Daniely, 2017), we study which representation, approximation, and resource properties persist or change under this coherent recursion.

Quantum evolution is linear in state amplitudes. This does not make a quantum model linear in its classical input: state preparation, tensor products, repeated encoding, and the Born rule can all generate nonlinear input dependence. It does mean that an intermediate activation cannot be inserted as an arbitrary scalar map while preserving coherence. Measuring each hidden layer and classically re-encoding its output is possible, but it generally discards coherences in the measured representation that a coherent recursion may otherwise retain.

This paper studies a direct alternative. A product of two reflections acts as a rotation in a two-dimensional invariant subspace, as in amplitude amplification (Brassard et al., 2002). We specialize this standard geometry to reflections about an encoded or current state and a trainable projector. Repeating the pair generates a Chebyshev polynomial of a projector overlap. Trainable reference states therefore play a role analogous to weights, their projector expectations supply overlap pre-activations, and repeated reflection pairs supply exact bounded response functions. An addressed bank of these responses, followed by an affine readout, is a quantum overlap analogue of a perceptron layer. The analogy is structural rather than literal: the pre-activation is a fidelity or reduced-state projector expectation, not an unconstrained signed affine form.

The construction is fully coherent between hidden layers: the hidden recursion requires no intermediate measurement, while parameter optimization and the terminal loss remain classical. Its defining feature is a neuron-level, exactly identifiable Chebyshev response and a recursive trainable comparison with the current quantum representation. Its relation to coherent data-reuploading circuits is discussed in Section 12.

The paper proceeds from one layer to coherent depth.

1. We define the projector-overlap unit and prove that alternating reflections implement the exact response $T_d(2p-1)$.
2. We characterize the complete idealized one-layer degree-D hypothesis space, including its tensor-power representation, dimension, strict hierarchy, kernel, restricted-domain capacity, interpolation properties, finite separation results, and Fourier structure.
3. We establish scalar Chebyshev approximation bounds, Universal Approximation Theorems for scalar and finite-output targets, and norm- and margin-sensitive generalization bounds.
4. We introduce a fully coherent recursive architecture. We prove layerwise preservation of the activation law and universal approximation, derive an unrolled process-feature representation, and identify an exact Chebyshev-composition subarchitecture.
5. We analyse the resulting depth–degree and resource tradeoffs and provide a simulator proof of concept on three eight-class image datasets.

The resulting theory separates three complementary questions: which functions the architecture represents, how those functions can be learned under statistical constraints, and what resources are required by a concrete circuit implementation.

2 Relation to prior work

2.1 Fixed features, kernels, and data re-uploading

Many supervised variational models with a fixed quantum feature state have the explicit form

$$f_\theta(x) = \mathrm{Tr}[\rho_x M_\theta]. \tag{1}$$

They are linear in ρ_x, even when ρ_x is a nonlinear function of the raw input. The corresponding Hilbert–Schmidt kernel is $K_1(x, x') = \mathrm{Tr}(\rho_x \rho_{x'})$. The feature-space and kernel perspectives of Schuld and Killoran (2019); Schuld (2021) make this structure explicit. Replacing a restricted variational observable family by an enlarged kernel class may

lower the optimized regularized empirical objective. This enlarged-feature perspective compares hypothesis classes, while the corresponding circuit realizations determine their operational resource requirements.

Repeated data encodings need a larger description. The Fourier spectrum is controlled by encoding-generator eigenvalue differences (Schuld et al., 2021), and data re-uploading can be universal even with few qubits (Pérez-Salinas et al., 2020). Jerbi et al. (2023) show that explicit, implicit, and re-uploading models can all be represented as linear quantum models after a suitable feature enlargement, while their qubit and sample requirements can differ substantially, as can operational resources not captured by the formal lift. Our tensor-power and process lifts give architecture-specific realizations of that enlarged-linear principle.

2.2 Coherent quantum networks and activation proposals

Prior work includes dissipative deep quantum networks, coherent feed-forward circuits, quantum-neuron proposals, and activation constructions (Beer et al., 2020; Tacchino et al., 2020; Singh et al., 2024; Cao et al., 2017; Maronese et al., 2022). Chebyshev-based quantum learning architectures also predate the present work. A recent preprint develops a hybrid quantum Chebyshev-polynomial network (Li et al., 2024).

Particularly close in motivation is QKAN, which applies trainable Chebyshev expansions to entries or eigenvalues of block-encoded diagonal operators by quantum singular-value transformation. Its Chebyshev implementation itself alternates signal-unitary and inverse calls with reflections (Ivashkov et al., 2026; Gilyén et al., 2019). Thus Chebyshev activations and alternating-reflection polynomial synthesis both have clear precedents. The contribution studied here is the encoded/current-state–projector overlap unit, its exact symmetric-tensor function-space characterization, and its coherent reduced-state recursion. QKAN and the present model differ in their signal variable, trainable geometry, access assumptions, recursion, and resource accounting.

Alternating reflections are fundamental in amplitude amplification and closely related to the two-dimensional signal transformations underlying quantum signal processing (Brassard et al., 2002; Low and Chuang, 2017). Repeating identical, unphased reflection pairs produces the Chebyshev sequence used here, while general phase-modulated quantum signal processing and singular-value transformation provide broader polynomial families (Gilyén et al., 2019). In the QFNN, trainable references and the terminal affine combination provide the learned feature dictionary and coefficients.

2.3 Classical depth as motivation

Classical depth-separation theorems show that compositional targets can admit efficient deep representations while requiring much wider shallow networks under specified activation, weight, distribution, and error assumptions (Telgarsky, 2016; Eldan and Shamir, 2016; Daniely, 2017). These results motivate the coherent recursion developed here. Because a QFNN has its own branch, oracle, gate, ancilla, measurement, and post-processing resources, Section 11 formulates the corresponding depth question directly for this architecture.

3 The one-layer quantum overlap perceptron

3.1 Data access and reference reflections

Let X be the input domain and $\mathcal{H} \cong \mathbb{C}^n$ the work Hilbert space. The encoding is a normalized pure state

$$\rho_x = |\phi_x\rangle\langle\phi_x|\,, \qquad |\phi_x\rangle = U_x\,|0\rangle\,. \tag{2}$$

The circuit assumes coherent access to both U_x and $U_x^\dagger$. This is stronger than receiving only copies of an unknown ρ_x; the latter model is not analysed. Generic state preparation can itself be expensive (Iten et al., 2016), so data-oracle calls and the compilation cost of U_x must be reported separately.

For a unit vector $|v\rangle$, define the Householder reflection

$$R_v = \mathbb{I} - 2\,|v\rangle\langle v|\,. \tag{3}$$

If $|v\rangle = U_v\,|0\rangle$, then $R_v = U_v(\mathbb{I} - 2\,|0\rangle\langle 0|)U_v^\dagger$. For an orthogonal projector $Q = Q^\dagger = Q^2$, define

$$R_Q = \mathbb{I} - 2Q. \tag{4}$$

3.2 A neuron-level analogy

A reference $|\chi_j\rangle$ defines $Q_j = |\chi_j\rangle\langle\chi_j|$ and the overlap pre-activation

$$F_j(x) = \mathrm{Tr}(\rho_x Q_j) = |\langle\chi_j|\phi_x\rangle|^2. \tag{5}$$

For reflection degree d_j, branch j applies

$$V_j^{(d_j)} = (R_{\phi_x} R_{Q_j})^{d_j}. \tag{6}$$

The centered branch response will be shown to equal

$$\alpha_{j,d_j}(x) = T_{d_j}(2F_j(x) - 1). \tag{7}$$

An address register can prepare and apply many branches coherently, and a terminal affine head forms

$$f(x) = b + \sum_{j=1}^{R} a_j \alpha_{j,d_j}(x). \tag{8}$$

The analogy with a classical perceptron is therefore

$$\text{reference state} \longrightarrow \text{overlap aggregation},$$
$$\text{Chebyshev response} \longrightarrow \text{affine readout}.$$

Unlike a classical perceptron, F_j is phase-insensitive and belongs to $[0, 1]$, while the centered quantity $2F_j - 1$ belongs to $[-1, 1]$. The model is consequently closer to a learned overlap or polynomial-feature unit than to an arbitrary signed affine neuron.

For C classes, the final centered features may be calibrated by

$$z_c(x) = b_c + \sum_j W_{cj}\alpha_j(x), \qquad c = 1, \ldots, C, \tag{9}$$

or by the parameter-light form $z_c = \beta\alpha_c + b_c$ when the response bank is class indexed. For classification, the logits are converted to predictive probabilities by softmax and trained with cross-entropy. Appendix B states the corresponding experimental protocol.

The affine calibration is optional at the operational level. One may use the class-indexed centered responses α_k directly as uncalibrated logits and apply a parameter-free softmax. Alternatively, whenever $\Pr(c = 0) > 0$, the normalized terminal probabilities

$$\Pr(A = k \mid c = 0) = \frac{w_k(1 + \alpha_k)}{\sum_r w_r(1 + \alpha_r)}$$

define class probabilities without a learned head. The affine readout is, however, part of the idealized function class in the span and density results below: those theorems require some mechanism for arbitrary finite real linear combinations and constants, whether implemented as a classical affine readout or by an equivalent terminal observable. This unrestricted readout is more general than the parameter-sharing calibration $z_k = \beta\alpha_k + b_k$ used in the numerical classifier, which has nine parameters for eight classes.

Unless stated otherwise, the function-space, pseudo-dimension, and Rademacher-complexity results below concern one real-valued score. A C-class model consists of C such scores coupled through Eq. (9). The approximation theorem has the immediate vector-valued extension stated in Corollary 7.5; the binary margin illustration in Section 9 instead uses labels in $\{-1, +1\}$ and is not presented as a multiclass bound.

3.3 Coherent extraction of the centered response

An interference qubit selects the identity or the addressed branch block. After a final Hadamard, if the address probability is $w_j > 0$,

$$p(j, 0) = w_j \frac{1 + \alpha_j}{2}, \qquad p(j, 1) = w_j \frac{1 - \alpha_j}{2}, \tag{10}$$

and hence

$$\alpha_j = \frac{p(j, 0) - p(j, 1)}{p(j, 0) + p(j, 1)}. \tag{11}$$

For a uniform R-branch address state this reduces to $\alpha_j = R[p(j, 0) - p(j, 1)]$. Thus the activation is the signed interference contrast between the two outcomes; either probability alone also contains the branch weight and a constant offset.

4 Alternating reflections give exact Chebyshev responses

The following projector-level identity is the mathematical core of the architecture. Its arbitrary-projector form will later cover reduced-state units in deep layers.

Theorem 4.1 (Projector-reflection Chebyshev identity) *Let $|\psi\rangle$ be a unit vector, let $P = |\psi\rangle\langle\psi|$, and let Q be any orthogonal projector. Define*

$$p = \langle\psi| Q |\psi\rangle, \qquad z = 2p - 1, \tag{12}$$

and $R_\psi = \mathbb{I} - 2P$, $R_Q = \mathbb{I} - 2Q$. For every $d \in \mathbb{N}_0$,

$$\langle\psi| (R_\psi R_Q)^d |\psi\rangle = T_d(z), \tag{13}$$

where $T_d(\cos\vartheta) = \cos(d\vartheta)$ is the Chebyshev polynomial of the first kind.

Proof. The cases $p = 0$ and $p = 1$ give $T_d(-1) = (-1)^d$ and $T_d(1) = 1$ directly. For $0 < p < 1$, set

$$|q\rangle = \frac{Q|\psi\rangle}{\sqrt{p}}, \qquad |q^\perp\rangle = \frac{(\mathbb{I} - Q)|\psi\rangle}{\sqrt{1-p}}. \tag{14}$$

Their span is invariant under both reflections. In the ordered basis $(|q\rangle, |q^\perp\rangle)$, the product $A = R_\psi R_Q$ is

$$A = \begin{pmatrix} z & -2\sqrt{p(1-p)} \\ 2\sqrt{p(1-p)} & z \end{pmatrix}. \tag{15}$$

Thus $\det A = 1$, $\operatorname{Tr} A = 2z$, and Cayley–Hamilton yields $A^{d+1} = 2zA^d - A^{d-1}$. With $a_d = \langle\psi| A^d |\psi\rangle$, one has $a_0 = 1$, $a_1 = z$, and

$$a_{d+1} = 2z a_d - a_{d-1}. \tag{16}$$

These are precisely the initial values and recurrence of $T_d(z)$. □

Corollary 4.2 (One-layer activation) *Every branch in Section 3 satisfies*

$$\langle\phi_x| V_j^{(d_j)} |\phi_x\rangle = T_{d_j}\big(2F_j(x) - 1\big), \tag{17}$$

and its response lies in $[-1, 1]$.

Proof. Apply Theorem 4.1 with $Q = Q_j$ and $|\psi\rangle = |\phi_x\rangle$. Since $F_j(x) \in [0, 1]$, the centered argument belongs to $[-1, 1]$, on which $|T_d| \le 1$. □

5 One-layer separation and tensor-power linearization

5.1 What is separated from a single-copy model

For a pure reference projector $Q = |\chi\rangle\langle\chi|$, define

$$\alpha_{Q,d}(\rho) = T_d(2\operatorname{Tr}(\rho Q) - 1). \tag{18}$$

The next result compares function classes on the full pure-state domain.

Theorem 5.1 (Separation from single-copy pure-state readouts) *Let $\dim\mathcal{H} \ge 2$, let $Q = |\chi\rangle\langle\chi|$, and let $d \ge 2$. There are no $M \in \mathrm{Herm}(\mathcal{H})$ and $b \in \mathbb{R}$ such that*

$$T_d\Big(2|\langle\chi|\psi\rangle|^2 - 1\Big) = \langle\psi| M |\psi\rangle + b \tag{19}$$

for every unit vector $|\psi\rangle \in \mathcal{H}$.

Proof. Absorb b into M as $b\mathbb{I}$. The left side is invariant under every unitary that fixes $|\chi\rangle$ and acts arbitrarily on $|\chi\rangle^\perp$. If Eq. (19) held for all pure states, equality of quadratic forms and polarization would make M commute with this unitary group. Hence

$$M = aQ + c(\mathbb{I} - Q) \tag{20}$$

for real a, c. Its expectation is $c + (a - c)|\langle\chi|\psi\rangle|^2$, an affine function of the fidelity. The left side is a degree-d polynomial in that fidelity with nonzero leading coefficient. Since the fidelity covers $[0, 1]$, equality is impossible for $d \ge 2$. □

Remark 5.2 (Comparator class) The comparator is the fixed-encoding single-copy class $\rho \mapsto \operatorname{Tr}(M\rho) + b$, with any trainable unitary absorbed into M. Multi-copy observables and data-reuploading circuits define larger comparator classes. Proposition 6.10 later gives an explicit finite sample on which the single-copy class is also separated.

5.2 Exact tensor-power lift

Theorem 5.3 (Tensor-power representation of one unit) *Write*

$$T_d(2t - 1) = \sum_{k=0}^{d} c_{d,k} t^k. \tag{21}$$

Then for every density operator ρ and rank-one projector Q,

$$\alpha_{Q,d}(\rho) = \sum_{k=0}^{d} c_{d,k} \operatorname{Tr}\Big(\rho^{\otimes k} Q^{\otimes k}\Big), \tag{22}$$

where the $k = 0$ term is the scalar 1.

Proof. For $k \geq 1$, multiplicativity of trace over tensor products gives

$$\operatorname{Tr}\big(\rho^{\otimes k} Q^{\otimes k}\big) = \big(\operatorname{Tr}(\rho Q)\big)^k. \tag{23}$$

Substitute this identity into Eq. (21). □

For example,

$$\alpha_{Q,1} = 2F_Q - 1, \tag{24}$$
$$\alpha_{Q,2} = 8F_Q^2 - 8F_Q + 1, \tag{25}$$
$$\alpha_{Q,3} = 32F_Q^3 - 48F_Q^2 + 18F_Q - 1. \tag{26}$$

Thus a degree-d unit is linear on

$$\Phi_{\leq d}(\rho) = 1 \oplus \rho \oplus \cdots \oplus \rho^{\otimes d}. \tag{27}$$

This is a representation identity for the feature space; the reflection circuit realizes the corresponding polynomial dependence through repeated queries to the preparation oracle.

Remark 5.4 (Relation to kernel descriptions) Theorem 5.1 is relative to the original single-copy feature $\rho \mapsto \rho$. Equation (27) induces, for positive weights η_k, the kernel

$$K_{\leq d}^{(\eta)}(\rho, \sigma) = \sum_{k=0}^{d} \eta_k \big[\operatorname{Tr}(\rho\sigma)\big]^k. \tag{28}$$

Every finite-degree QFNN unit is a linear functional in this enlarged feature space. The construction is therefore compatible with the fixed feature kernel perspective of Schuld (2021) and the enlarged linear-model framework of Jerbi et al. (2023). The tensor lift identifies the corresponding polynomial-kernel class while retaining its specific alternating-reflection circuit realization.

6 Exact finite-degree expressivity on pure states

The tensor identity for one unit can be strengthened to an exact characterization of the complete idealized bounded-degree family. This section studies arbitrary pure references, arbitrary finite width, all degrees up to D, and an unrestricted real affine combination.

6.1 The symmetric tensor space

For $D \in \mathbb{N}$, let $P_{\mathrm{sym}}^{(D)}$ project onto $\operatorname{Sym}^D \mathcal{H}$, and write $\rho_\psi = |\psi\rangle\langle\psi|$. Define the real function space

$$\mathcal{V}_D = \big\{h_A : \quad h_A(\rho_\psi) = \operatorname{Tr}\Big(A\rho_\psi^{\otimes D}\Big), \quad A \in \operatorname{Herm}(\operatorname{Sym}^D \mathcal{H})\big\}. \tag{29}$$

For degree zero, use the conventions

$$\operatorname{Sym}^0 \mathcal{H} = \mathbb{C}, \qquad \mathcal{V}_0 = \mathbb{R}\mathbf{1}. \tag{30}$$

Lemma 6.1 (Coherent rank-one operators span the symmetric operator space) *For finite-dimensional $\mathcal{H}$,*

$$\operatorname{span}_{\mathbb{R}} \Big\{|v^{\otimes D}\rangle\langle v^{\otimes D}| : |v\rangle \in \mathcal{H}\Big\} = \operatorname{Herm}(\operatorname{Sym}^D \mathcal{H}). \tag{31}$$

Proof. Suppose $A \in \operatorname{Herm}(\operatorname{Sym}^D \mathcal{H})$ is Hilbert–Schmidt orthogonal to every $|v^{\otimes D}\rangle\langle v^{\otimes D}|$. In a symmetric occupation-number basis indexed by multi-indices μ, ν with $|\mu| = |\nu| = D$, this means

$$0 = \langle z|^{\otimes D} A\, |z^{\otimes D}\rangle = \sum_{|\mu|=|\nu|=D} a_{\mu\nu}\, \overline{z}^{\mu} z^{\nu} \tag{32}$$

for every $z \in \mathbb{C}^n$, with nonzero normalization constants absorbed into $a_{\mu\nu}$. The monomials $\overline{z}^{\mu} z^{\nu}$ are linearly independent as polynomials in the real and imaginary parts of z. Equivalently, applying the Wirtinger derivative $\partial_{\overline{z}}^{\mu} \partial_z^{\nu}$ at $z = 0$ isolates the coefficient $a_{\mu\nu}$. Hence every coefficient vanishes and $A = 0$. The orthogonal complement of the real span in Eq. (31) is therefore trivial. □

Theorem 6.2 (Exact degree-D QFNN function space) *On normalized pure states,*

$$\mathcal{V}_D = \operatorname{span}_{\mathbb{R}}\Big(\{1\} \cup \big\{\alpha_{Q,d} : Q \text{ pure}, \quad 1 \leq d \leq D\big\}\Big). \tag{33}$$

Proof. Let Q be pure and $k \leq D$. Since $|\psi\rangle^{\otimes D}$ is symmetric and normalized,

$$\big(\operatorname{Tr}(\rho_\psi Q)\big)^k = \langle\psi|^{\otimes D} P_{\mathrm{sym}}^{(D)}$$

$$\times \left(Q^{\otimes k} \otimes \mathbb{I}^{\otimes(D-k)}\right) P_{\mathrm{sym}}^{(D)} |\psi\rangle^{\otimes D}. \tag{34}$$

Every $\alpha_{Q,d}$ with $d \le D$ is a polynomial in $\mathrm{Tr}(\rho_\psi Q)$, so the right-hand side of Eq. (33) is contained in $\mathcal{V}_D$.

Conversely, Lemma 6.1 writes every $A \in \mathrm{Herm}(\mathrm{Sym}^D \mathcal{H})$ as a finite real linear combination of $Q^{\otimes D}$ with pure Q. Thus every function in $\mathcal{V}_D$ is a finite linear combination of $[\mathrm{Tr}(\rho_\psi Q)]^D$. Since t^D belongs to the span of $T_0(2t-1), \ldots, T_D(2t-1)$, those functions are finite combinations of QFNN atoms of degrees at most D. □

Corollary 6.3 (Dimension and strict hierarchy) *Let $n = \dim \mathcal{H}$ and*

$$s_D = \dim \mathrm{Sym}^D \mathcal{H} = \binom{n+D-1}{D}. \tag{35}$$

Then

$$\dim_{\mathbb{R}} \mathcal{V}_D = s_D^2 = \binom{n+D-1}{D}^2. \tag{36}$$

For $n \ge 2$,

$$\mathcal{V}_0 \subsetneq \mathcal{V}_1 \subsetneq \mathcal{V}_2 \subsetneq \cdots. \tag{37}$$

Proof. Hermitian operators on an s_D-dimensional complex space form a real vector space of dimension s_D^2. The linear map

$$A \longmapsto \left[\rho_\psi \longmapsto \mathrm{Tr}(A\rho_\psi^{\otimes D})\right] \tag{38}$$

is injective by the polynomial argument in Lemma 6.1. Lower-degree functions embed by Eq. (34), and $s_D > s_{D-1}$ for $n \ge 2$, which makes every inclusion strict. □

At fixed D and large n, the ambient dimension satisfies $\dim \mathcal{V}_D \sim n^{2D}/(D!)^2$; at fixed n and large D, it satisfies $\dim \mathcal{V}_D \sim D^{2(n-1)}/((n-1)!)^2$. These formulas quantify the complete idealized space; the reachable subset of a parameterized reference ansatz is determined by its own geometry.

Corollary 6.4 (Homogeneous polynomial fidelity kernel) *Set $\Phi_D(\rho) = \rho^{\otimes D}$. Its Hilbert–Schmidt kernel is*

$$K_D(\rho,\sigma) = \mathrm{Tr}\left(\rho^{\otimes D}\sigma^{\otimes D}\right) = \left[\mathrm{Tr}(\rho\sigma)\right]^D. \tag{39}$$

For pure states,

$$K_D(\psi,\phi) = |\langle\psi|\phi\rangle|^{2D}. \tag{40}$$

On the normalized pure-state domain, the unrestricted real function space generated by this kernel is $\mathcal{V}_D$.

Remark 6.5 (Function space versus RKHS norm) On normalized pure states, the homogeneous kernel in Eq. (39) and positively weighted inhomogeneous kernels of the form Eq. (28) span the same degree-at-most-D functions, but generally assign different norms. In particular, lower-degree terms and constants can have degree-dependent homogeneous-kernel norms. Norm-controlled generalization comparisons must therefore state which kernel and normalization they use.

6.2 Capacity on the encoded domain

Definition 6.6 (Order-D feature rank) For an encoded domain X, define

$$r_D(X) = \dim_{\mathbb{R}} \mathrm{span}\left\{\rho_x^{\otimes D} : x \in X\right\}. \tag{41}$$

For $D = 0$ and nonempty X, use $\rho_x^{\otimes 0} = 1$ and $r_0(X) = 1$.

Theorem 6.7 (Exact restricted-domain dimension and pseudo-dimension) *For the unrestricted linear class $\mathcal{V}_D$ restricted to X,*

$$\dim(\mathcal{V}_D|_X) = \mathrm{Pdim}(\mathcal{V}_D|_X) = r_D(X). \tag{42}$$

For a finite sample $S = \{x_1, \ldots, x_m\}$,

$$r_D(S) = \mathrm{rank}\left[K_D(x_i,x_j)\right]_{i,j=1}^m, \tag{43}$$

and

$$r_{D-1}(S) \le r_D(S) \le \min\left\{m, \binom{n+D-1}{D}^2\right\}. \tag{44}$$

Proof. Restricting all linear functionals of the feature $\rho_x^{\otimes D}$ gives the dual of its real feature span, hence the first dimension equality. Any real r-dimensional vector space of functions has pseudo-dimension at most r. Conversely, choose $r = r_D(X)$ points whose evaluation vectors are linearly independent. The resulting evaluation map onto $\mathbb{R}^r$ is surjective, so arbitrary values, and hence arbitrary signs at zero thresholds, can be interpolated on those points. Thus the pseudo-dimension is at least r.

The rank of a Gram matrix equals the dimension of the span of its feature vectors, proving Eq. (43). Partial trace maps $\rho_x^{\otimes D}$ linearly to $\rho_x^{\otimes(D-1)}$, so feature rank cannot decrease with D. The remaining bounds are immediate. □

Remark 6.8 (Full space versus a fixed finite bank) Equation (42) describes the complete idealized linear space $\mathcal{V}_D|_X$. For a fixed dictionary of R atoms $\phi_1, \ldots, \phi_R$, the affine class

$$b + \sum_{j=1}^{R} a_j \phi_j$$

has pseudo-dimension $\dim \operatorname{span}\{1, \phi_1, \ldots, \phi_R\} \leq R+1$. When reference states vary across hypotheses, the dictionary itself is trainable and the resulting class is a union of such subspaces. Conversely, a fixed bank spanning all of $\mathcal{V}_D$ requires at least $\dim \mathcal{V}_D - 1$ nonconstant atoms, although individual targets may admit much smaller representations.

6.3 Finite-sample interpolation and an explicit separation

Corollary 6.9 (Finite-sample interpolation) *Let $\rho_1, \ldots, \rho_m$ be distinct pure states. Every real assignment $y_1, \ldots, y_m$ is interpolated by some function in $\mathcal{V}_{m-1}$ and hence by a finite variable-degree QFNN bank of maximum degree $m-1$.*

Proof. If $m = 1$, the constant function $f(\rho) = y_1$ belongs to $\mathcal{V}_0$. Assume henceforth that $m \geq 2$. For $i \neq j$, let $A_{ij} = \rho_i - \rho_j$ and $\Delta_{ij} = \operatorname{Tr}[(\rho_i - \rho_j)^2] > 0$. Define

$$g_{ij}(\rho) = \frac{\operatorname{Tr}(A_{ij}\rho) - \operatorname{Tr}(A_{ij}\rho_j)}{\Delta_{ij}}. \tag{45}$$

Then $g_{ij}(\rho_i) = 1$ and $g_{ij}(\rho_j) = 0$. The polynomial

$$p_i(\rho) = \prod_{j \neq i} g_{ij}(\rho) \tag{46}$$

has degree at most $m-1$ and satisfies $p_i(\rho_k) = \delta_{ik}$. Therefore $f(\rho) = \sum_i y_i p_i(\rho)$ interpolates the labels. Products of the affine expectations can be written as a symmetrized order-$(m-1)$ tensor expectation: because $\operatorname{Tr}\rho = 1$, each $g_{ij}(\rho) = \operatorname{Tr}(B_{ij}\rho)$ for a Hermitian B_{ij}, and

$$p_i(\rho) = \operatorname{Tr}\left[\rho^{\otimes(m-1)} \bigotimes_{j \neq i} B_{ij}\right]. \tag{47}$$

Sandwiching the tensor-product operator by $P_{\mathrm{sym}}^{(m-1)}$ does not change its expectation on a pure tensor power. Thus $f \in \mathcal{V}_{m-1}$. Apply Theorem 6.2. □

Thus the idealized family has exact finite-sample interpolation capacity by degree $m-1$.

Proposition 6.10 (Finite alternating-label separation) *For $D \geq 2$, consider the equatorial qubit states*

$$|\psi_{\theta_k}\rangle = \frac{|0\rangle + e^{i\theta_k}|1\rangle}{\sqrt{2}}, \qquad \theta_k = \frac{k\pi}{D}, \quad k = 0, \ldots, 2D-1. \tag{48}$$

with labels $y_k = (-1)^k$ and the uniform distribution. A single degree-D QFNN unit classifies every point with margin one. Every affine single-copy observable score f has nonpositive-margin risk

$$\frac{1}{2D}\#\{k : y_k f(\theta_k) \leq 0\} \geq \frac{D-1}{2D}. \tag{49}$$

Proof. For $Q = |+\rangle\langle+|$,

$$2\operatorname{Tr}(\rho_\theta Q) - 1 = \cos\theta, \tag{50}$$

so

$$\alpha_{Q,D}(\rho_{\theta_k}) = T_D(\cos\theta_k) = \cos(D\theta_k) = (-1)^k. \tag{51}$$

An arbitrary qubit observable plus bias restricts on the equator to

$$f(\theta) = c + a\cos\theta + b\sin\theta. \tag{52}$$

If it is identically zero, all $2D$ points are nonpositive-margin errors. Otherwise its positive points form at most one cyclic block. Complete the prediction to a binary cyclic sequence by declaring $f \leq 0$ negative. This completion has at most two transitions and can only increase the correct count relative to the strict-margin score because sampled zeros are errors in the proposition. Against a length $2D$ alternating label sequence, any binary cyclic sequence with at most two transitions has at most $D+1$ correct entries. Thus the original score makes at least $D-1$ nonpositive-margin errors. □

Remark 6.11 For $D = 2$, Proposition 6.10 is a four-state XOR-type separation from the fixed-encoding single-copy affine-observable class. Multi-copy, polynomial-kernel, and data-reuploading models belong to larger comparator classes.

7 Approximation theory and Universal Approximation Theorems

7.1 Scalar overlap responses and the role of degree

For a scalar response $g : [0,1] \to \mathbb{R}$, define the best degree-D error

$$E_D(g) = \inf_{\substack{p \in \mathbb{R}[F] \\ \deg p \le D}} \sup_{0 \le F \le 1} |g(F) - p(F)| . \tag{53}$$

Equivalently, write $p(F) = \sum_{d=0}^{D} a_d T_d(2F-1)$, because the shifted Chebyshev polynomials form a basis of $\mathbb{R}[F]_{\le D}$. The quantity is independent of a particular reference Q: it is a uniform envelope over the whole attainable overlap interval. Because the degree spaces are nested,

$$E_{D+1}(g) \le E_D(g). \tag{54}$$

Thus increasing degree cannot worsen the best attainable scalar approximation. The inequality need not be strict if the target is already represented at lower degree.

Theorem 7.1 (Chebyshev approximation of an overlap response) *Let $g \in C([0,1], \mathbb{R})$. Then*

$$\lim_{D \to \infty} E_D(g) = 0. \tag{55}$$

If $\widetilde{g}(z) = g((z+1)/2)$ extends analytically to the open Bernstein ellipse E_ϱ, is continuous on its boundary, and satisfies $|\widetilde{g}(z)| \le M$ there for some $\varrho > 1$, then

$$E_D(g) \le \frac{2M\varrho^{-D}}{\varrho - 1}. \tag{56}$$

Proof. $T_0, \dots, T_D$ form a basis of the real polynomials of degree at most D. Polynomial density in $C([-1,1])$ gives the first claim after the affine change of variables $z = 2F - 1$. Under the analyticity assumption, the Chebyshev coefficients obey $|a_d| \le 2M\varrho^{-d}$. Summing the tail from $d = D+1$ yields Eq. (56) (Trefethen, 2019). □

For analytic scalar targets, Eq. (56) certifies a geometric decrease of the approximation-error bound with reflection degree D. For general continuous targets, Theorem 7.1 guarantees convergence without prescribing a universal rate.

7.2 Universal Approximation Theorem for the idealized one-layer family

In contrast to scalar approximation through one fixed overlap, the density result allows a finite bank whose width, references, and degrees depend on the requested accuracy. Its elements have the explicit form

$$f(x) = b + \sum_{j=1}^{R} a_j T_{d_j}(2 \operatorname{Tr}(\rho_x Q_j) - 1) , \tag{57}$$

with arbitrary finite R, pure references Q_j, and nonnegative integer degrees d_j. The uniform closure below is taken over the union of all such finite banks. This is the precise analogue of varying the finite hidden width in a classical one-hidden-layer approximation theorem.

Theorem 7.2 (Universal Approximation Theorem for one-layer pure-state QFNNs) *Let X be compact Hausdorff and let $\rho : X \to \mathsf{D}(\mathcal{H})$ be continuous, injective, and pure-valued. The uniform closure of the union of one-layer QFNN models over arbitrary finite widths, arbitrary pure references, arbitrary reflection degrees, and an affine readout is $C(X, \mathbb{R})$.*

Proof. The map ρ is a homeomorphism from X onto its compact image $K = \rho(X)$. By Theorem 6.2 for $D \ge 1$, together with the convention $\mathcal{V}_0 = \mathbb{R}\mathbf{1}$ and the affine bias, the one-layer family contains every function

$$h_A(\sigma) = \operatorname{Tr}(A\sigma^{\otimes D}), \qquad D \in \mathbb{N}_0, \qquad A \in \operatorname{Herm}(\operatorname{Sym}^D \mathcal{H}). \tag{58}$$

These functions and constants form a real unital algebra on K: for pure σ,

$$h_A(\sigma) h_B(\sigma) = \operatorname{Tr}\left[\sigma^{\otimes(D+E)}(A \otimes B)\right], \tag{59}$$

with the operator restricted or symmetrized on $\operatorname{Sym}^{D+E} \mathcal{H}$. They separate points because Hermitian linear expectations already separate distinct density operators. The real Stone–Weierstrass theorem makes the algebra uniformly dense in $C(K, \mathbb{R})$ (Stone, 1948). Pullback by ρ completes the proof. □

Remark 7.3 (Idealized family and state domain) The theorem concerns the union over arbitrary finite widths and degrees, unrestricted pure references, and affine readouts. Purity enters through the exact function-space characterization in Theorem 6.2. Extending that characterization to mixed-state encodings requires a separate analysis of the attainable polynomial algebra.

Corollary 7.4 (L^p density) *Under the continuous, injective, pure-valued encoding assumptions of Theorem 7.2, if X is compact metric and μ is a finite regular Borel measure, the same family is dense in $L^p(X,\mu)$ for every $1 \le p < \infty$.*

Proof. Continuous functions are dense in $L^p(X,\mu)$, and uniform convergence implies L^p convergence on a finite measure space. Apply Theorem 7.2 in a two-stage approximation. □

Corollary 7.5 (Finite-output Universal Approximation Theorem) *Under the assumptions of Theorem 7.2, for every finite output dimension C, the union of finite QFNN banks with the unrestricted matrix head in Eq. (9) is uniformly dense in $C(X,\mathbb{R}^C)$. If, in addition, X is compact metric and μ is a finite regular Borel measure, the same family is dense in $L^p(X,\mu;\mathbb{R}^C)$ for every $1 \le p < \infty$.*

Proof. Apply Theorem 7.2, or Corollary 7.4, to each of the C coordinate functions. The union of the resulting finite scalar banks is still finite, and the rows of W together with the bias vector select the required coordinatewise linear combinations. □

These density statements concern real score vectors. In classification, one applies the discontinuous arg max only after constructing the scores; uniform score approximation transfers labels away from zero multiclass margin, while L^p density alone does not control tie sets.

8 Amplitude and Fourier signal views

8.1 Amplitude encoding

Let $a(x) \in \mathbb{C}^n$ be normalized and $|\phi_x\rangle = \sum_r a_r(x)\,|r\rangle$. For $|\chi\rangle = \sum_r \chi_r\,|r\rangle$,

$$F_\chi(x) = \left|\sum_r \chi_r^* a_r(x)\right|^2 = \sum_{r,s} \chi_r^* \chi_s a_r(x) a_s(x)^*. \tag{60}$$

Consequently $T_d(2F_\chi - 1)$ has bidegree at most (d,d) in (a, a^*). For real normalized amplitudes and references, it has degree at most $2d$. If a raw nonzero vector is encoded as $a(x) = x/\|x\|_2$, the response is generally rational in the raw coordinates; it is polynomial only on the normalized sphere or in the normalized amplitudes.

8.2 Fourier support

Suppose a periodic parameterization gives

$$g_\chi(x) = \langle \chi | \phi_x \rangle = \sum_{\omega \in \Omega} a_{\chi,\omega} e^{i\omega \cdot x}, \tag{61}$$

with finite Ω. For sets A, B, write $A - B = \{a - b : a \in A, b \in B\}$ and let mA be the m-fold Minkowski sum, with $0A = \{0\}$.

Proposition 8.1 (Fourier support of a Chebyshev overlap) *Under Eq. (61),*

$$\operatorname{supp} F_\chi \subseteq \Omega - \Omega \tag{62}$$

and

$$\operatorname{supp} \alpha_{\chi,d} \subseteq \bigcup_{m=0}^{d} m(\Omega - \Omega). \tag{63}$$

Proof. Expanding $|g_\chi|^2$ pairs frequencies as $\omega - \omega'$. The degree-d response is a polynomial in F_χ, and multiplication of Fourier series adds frequencies. □

This recovers the expected Minkowski-sum growth from encoding and data-reuploading Fourier theory (Schuld et al., 2021).

8.3 Coefficient rank

Support alone can substantially overstate functional freedom. Assume x lies on a standard torus

(or another periodic domain on which the displayed characters are linearly independent), so Fourier coefficients are unique. Suppose the homogeneous lifted feature has the finite expansion

$$\rho_x^{\otimes D} = \sum_{\lambda \in \Lambda_D} e^{i\lambda \cdot x} B_\lambda. \tag{64}$$

The operators B_λ are understood as compressed to $\mathrm{Sym}^D \mathcal{H}$. Hermiticity and Fourier uniqueness imply $B_{-\lambda} = B_\lambda^\dagger$. After adjoining zero coefficients if necessary, take $\Lambda_D = -\Lambda_D$.

Proposition 8.2 (Fourier coefficient-rank characterization) *Choose one representative from every nonzero pair $\{\lambda, -\lambda\}$. The real dimension of the Fourier coefficient vectors generated by the full class $\mathcal{V}_D$ equals*

$$\dim_{\mathbb{R}} \mathrm{span}\Big\{ B_0,\ \tfrac{1}{2}(B_\lambda + B_\lambda^\dagger),\ \tfrac{1}{2i}(B_\lambda - B_\lambda^\dagger) \Big\}, \tag{65}$$

and is at most

$$\binom{n+D-1}{D}^2. \tag{66}$$

Proof. For $f_A(x) = \mathrm{Tr}(A\rho_x^{\otimes D})$,

$$\widehat{f_A}(\lambda) = \mathrm{Tr}(AB_\lambda). \tag{67}$$

Its real and imaginary coefficient coordinates are the real linear functionals represented by the Hermitian real and imaginary parts displayed in Eq. (65). The rank of this coefficient map equals the real span rank of its representing operators and cannot exceed $\dim \mathcal{V}_D$. □

For an equatorial qubit with Bloch vectors $r(x) = (\cos x, \sin x, 0)$ and $u(\varphi) = (\cos\varphi, \sin\varphi, 0)$,

$$\alpha_{\varphi,d}(x) = T_d(\cos(x-\varphi)) = \cos(d(x-\varphi)). \tag{68}$$

Two phases for each $d \geq 1$ generate both $\cos(dx)$ and $\sin(dx)$. Degrees $0, \ldots, D$ therefore span exactly the $2D+1$ real trigonometric polynomials on this curve, even though the full qubit pure-state space has $\dim \mathcal{V}_D = (D+1)^2$. This illustrates both support saturation and restriction-induced rank collapse without saturating the ambient pure-state dimension.

9 Statistical capacity and generalization

Expressivity, approximation, and generalization are distinct. The strict hierarchy in Eq. (37) enlarges the unconstrained hypothesis space, while Eq. (54) decreases its best target-specific approximation error. Which effect dominates population risk depends on the target, data distribution, regularization, optimization, and selected solution.

Throughout this section, $\widehat{\mathfrak{R}}_S$ denotes the complexity of one scalar score class. The displayed margin corollary is stated for binary labels. Multiclass extensions follow through vector-contraction bounds and depend on the norm chosen for the matrix head in Eq. (9).

9.1 A norm-controlled lifted bound

For $B > 0$, define

$$\mathcal{F}_{D,B} = \big\{ f_A : f_A(x) = \mathrm{Tr}(A\rho_x^{\otimes D}),\ A \in \mathrm{Herm}(\mathrm{Sym}^D \mathcal{H}),\ \|A\|_{\mathrm{HS}} \leq B \big\}. \tag{69}$$

Any bias must be absorbed into this bounded operator or bounded separately.

Theorem 9.1 (Rademacher complexity of the lifted norm ball) *For a sample $S = (x_1, \ldots, x_m)$ of pure-state inputs,*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_{D,B}) \leq \frac{B}{\sqrt{m}}. \tag{70}$$

The same norm proof gives $B/\sqrt{m}$ for arbitrary density operators with feature $\rho_x^{\otimes D}$ and $A \in \mathrm{Herm}(\mathcal{H}^{\otimes D})$. The exact mixed-state QFNN span is not characterized here.

Proof. With independent Rademacher signs ε_i,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_{D,B}) = \frac{1}{m}\mathbb{E}_\varepsilon \sup_{\|A\|_{\mathrm{HS}} \leq B} \left\langle A, \sum_{i=1}^m \varepsilon_i \rho_{x_i}^{\otimes D} \right\rangle_{\mathrm{HS}} \tag{71}$$

$$= \frac{B}{m}\mathbb{E}_\varepsilon \left\| \sum_{i=1}^m \varepsilon_i \rho_{x_i}^{\otimes D} \right\|_{\mathrm{HS}} \tag{72}$$

$$\leq \frac{B}{m} \sqrt{\sum_{i=1}^m \left\| \rho_{x_i}^{\otimes D} \right\|_{\mathrm{HS}}^2} \leq \frac{B}{\sqrt{m}}. \tag{73}$$

The penultimate step uses Jensen's inequality and cancellation of cross terms under the signs. □

Standard Rademacher risk bounds imply that a bounded L-Lipschitz loss has a uniform generalization gap of order

$$O\left(\frac{LB}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{m}}\right) \tag{74}$$

with probability at least $1-\delta$ (Bartlett and Mendelson, 2002). The hidden constant in the concentration term depends on the numerical range bound of the loss. Boundedness, or assumptions that imply an appropriate tail bound, is necessary in addition to Lipschitz continuity.

Remark 9.2 (Where degree enters the bound) Equation (70) has the form $B/\sqrt{m}$ at fixed Hilbert–Schmidt radius. Degree dependence enters through the norm required to represent or approximate the target and through the attainable approximation error and margin. Higher-degree lifting is favorable when it improves approximation at a comparable norm or increases the margin.

9.2 Fixed-bank head and margins

Proposition 9.3 (Fixed-bank head complexity) *Fix R QFNN atoms independently of the sample and let $\alpha(x) = (\alpha_1(x), \dots, \alpha_R(x))$. For linear heads with $\|a\|_2 \le B$,*

$$\widehat{\mathfrak{R}}_S \le B\sqrt{\frac{R}{m}}. \tag{75}$$

For $\|a\|_1 \le A$,

$$\widehat{\mathfrak{R}}_S \le A\sqrt{\frac{2\log(2R)}{m}}. \tag{76}$$

A bounded bias can be included as an extra constant coordinate.

Proof. Corollary 4.2 gives $\|\alpha(x)\|_2 \le \sqrt{R}$, yielding Eq. (75) by Hilbert-space duality. For the ℓ_1 class, duality reduces the supremum to the largest of R bounded signed sums; the standard maximal inequality gives Eq. (76). □

For binary labels $y \in \{-1, +1\}$, applying a $1/\gamma$-Lipschitz ramp loss gives the usual schematic margin bound

$$\mathbb{P}(yf(x) \le 0) \le \widehat{\mathbb{P}}(y_i f(x_i) \le \gamma) + O\left(\frac{\widehat{\mathfrak{R}}_S}{\gamma} + \sqrt{\frac{\log(1/\delta)}{m}}\right). \tag{77}$$

The degree does not appear explicitly because every Chebyshev atom is bounded. If the reference states are learned on the same sample, the complexity of choosing the atom bank must also be included. Circuit-level covering bounds and gate-sharing results provide alternative controls (Caro et al., 2022).

9.3 Target alignment and kernel statistical dimension

Proposition 6.10 exhibits a target having zero degree-D risk and irreducible degree-one risk. Combining such an approximation separation with a norm or margin bound can yield a target-specific approximation–estimation advantage once the sample is large enough.

For a sample Gram matrix K_D and $\lambda > 0$, the regularized kernel statistical dimension

$$d_{\mathrm{eff},D}(\lambda) = \mathrm{Tr}\left[K_D(K_D + m\lambda I)^{-1}\right] \tag{78}$$

satisfies

$$d_{\mathrm{eff},D}(\lambda) \le \mathrm{rank}\, K_D = r_D(S). \tag{79}$$

It is distinct from the Fisher-information effective dimension used by Abbas et al. (2021). In the present pure-state setting it has an additional monotonicity property.

Proposition 9.4 (Pure-state kernel effective dimension is nondecreasing) *For a fixed sample of pure states and every $\lambda > 0$,*

$$d_{\mathrm{eff},D+1}(\lambda) \ge d_{\mathrm{eff},D}(\lambda). \tag{80}$$

Proof. Put $G = K_1$. Then $G \succeq 0$, $G_{ii} = 1$, and $K_{D+1} = G \circ K_D$, where $\circ$ is the Hadamard product. The Schur map $\Phi_G(X) = G \circ X$ is completely positive, unital, and trace preserving. By the Ky Fan variational principle, a unital trace-preserving positive map sends the eigenvalue vector of a Hermitian input to one majorized by the input vector.

Hence the eigenvalue vector of $K_{D+1} = \Phi_G(K_D)$ is majorized by that of K_D. For $a = m\lambda > 0$, the function $h_a(t) = t/(t+a)$ is concave on $[0, \infty)$. Karamata's inequality therefore gives

$$\sum_i h_a(\lambda_i(K_{D+1})) \geq \sum_i h_a(\lambda_i(K_D)), \qquad (81)$$

which is Eq. (80). □

Purity enters this proof through $G_{ii} = 1$. For mixed inputs the associated Schur map is not generally unital, so the monotonicity conclusion does not follow from the same majorization argument. This effective dimension provides a natural data-dependent statistic for future kernel-ridge and empirical QFNN studies.

Recent work motivates data-dependent and margin-based analyses beyond uniform parameter counts (Gil-Fuster et al., 2024; Hur and Park, 2025). The present results supply exact feature geometry for such studies.

10 Deep coherent QFNNs

The one-layer tensor picture identifies the exact finite-degree hypothesis space. We now use the same overlap primitive recursively. All address, interference, and work registers may be retained coherently, allowing later projectors to interrogate the hidden representation without intermediate measurement or classical re-encoding. Every such recursive comparison continues to obey the exact Chebyshev activation law.

10.1 Layer definition and retained coherences

Let the normalized incoming state to layer ℓ be $|\Phi_x^{(\ell-1)}\rangle$ on all registers retained from previous layers. For fixed preceding parameters $\theta_{<\ell}$, let $U_{\theta_{<\ell}}^{(\ell-1)}(x)$ denote the complete preparation through layer $\ell-1$, so that $U_{\theta_{<\ell}}^{(\ell-1)}(x)\,|0\rangle = |\Phi_x^{(\ell-1)}\rangle$. Define the corresponding reflection by

$$R_\Phi^{(\ell)} = \mathbb{I} - 2\,|\Phi_x^{(\ell-1)}\rangle\langle\Phi_x^{(\ell-1)}|\,. \qquad (82)$$

Because this entire preceding preparation, including the fixed trainable gates in layers $1, \ldots, \ell-1$, is known,

$$R_\Phi^{(\ell)} = U_{\theta_{<\ell}}^{(\ell-1)}(x)(\mathbb{I} - 2\,|0\rangle\langle 0|)U_{\theta_{<\ell}}^{(\ell-1)}(x)^\dagger. \qquad (83)$$

This implementability assumption is stronger than access to unknown copies of the incoming state and is accounted for in Section 11.

The current-state reflection acts on the complete retained pure state. By contrast, branch j may use a global reference projector or a projector $Q_{j,S}^{(\ell)}$ on a selected subsystem $S_{\ell-1}$. In the latter case, its full-space extension is

$$\widetilde{Q}_j^{(\ell)} = \mathbb{I}_{\overline{S}_{\ell-1}} \otimes Q_{j,S}^{(\ell)}. \qquad (84)$$

For a global reference, set $\widetilde{Q}_j^{(\ell)} = Q_j^{(\ell)}$. The branch applies

$$V_j^{(\ell)} = \left(R_\Phi^{(\ell)} R_{\widetilde{Q}_j^{(\ell)}}\right)^{d_j^{(\ell)}}. \qquad (85)$$

The local choice targets a reduced hidden representation, such as the address–interference registers used in the numerical architecture; a global reference can additionally access distinctions and correlations stored in the complete retained state. Let the address weights be a probability distribution, $w_j^{(\ell)} \geq 0$ and $\sum_j w_j^{(\ell)} = 1$. The post-Hadamard layer state can be written as

$$|\Phi_x^{(\ell)}\rangle = \sum_j \sqrt{w_j^{(\ell)}}\,|j\rangle \left[|0\rangle \frac{\mathbb{I} + V_j^{(\ell)}}{2} |\Phi_x^{(\ell-1)}\rangle + |1\rangle \frac{\mathbb{I} - V_j^{(\ell)}}{2} |\Phi_x^{(\ell-1)}\rangle \right]. \qquad (86)$$

The address and interference registers need not be measured. Thus off-diagonal address terms and coherences between the two interference arms remain available to later projectors. Tracing them out is an optional design choice, not part of the mathematical definition.

For the local case, let $\rho_{S_{\ell-1}}^{(\ell-1)}(x)$ be the reduced incoming state. Then

$$\begin{aligned} p_j^{(\ell)}(x) &= \langle\Phi_x^{(\ell-1)}|\,\widetilde{Q}_j^{(\ell)}\,|\Phi_x^{(\ell-1)}\rangle \\ &= \mathrm{Tr}\left[\rho_{S_{\ell-1}}^{(\ell-1)}(x) Q_{j,S}^{(\ell)}\right]. \end{aligned} \qquad (87)$$

Corollary 10.1 (Layerwise exact activation) *For every layer and branch defined above, the centered interference response is*

$$\alpha_j^{(\ell)}(x) = T_{d_j^{(\ell)}}\left(2p_j^{(\ell)}(x) - 1\right). \tag{88}$$

Proof. Apply Theorem 4.1 to the full incoming pure state and to $\widetilde{Q}_j^{(\ell)}$. Equation (87) rewrites the same expectation on the selected reduced state in the local case. □

This corollary is the exact sense in which the activation law survives depth. In a generic layer, $p_j^{(\ell)}$ depends on the full coherently generated state, so the next layer acts on the retained representation rather than only on the separately reported scalar responses from layer $\ell - 1$. Figure 2 shows the distinction between a local reference reflection and the global reflection about the retained incoming state.

10.2 An exactly compositional scalar-routing subarchitecture

There is a nontrivial subarchitecture in which literal scalar composition is exact. It provides a constructive depth result inside the coherent QFNN.

Theorem 10.2 (Exact Chebyshev routing through coherent layers) *Choose one branch in every layer. Let*

$$z_0 = 2\langle\phi_x| Q_1 |\phi_x\rangle - 1, \tag{89}$$

and let the first layer have degree d_1. For each later layer $\ell \geq 2$, choose the local projector

$$Q_\ell = |0\rangle\langle 0|_{c_{\ell-1}} \otimes \mathbb{I}, \tag{90}$$

where $c_{\ell-1}$ is the previous layer's interference qubit after its final Hadamard. Then the centered responses satisfy

$$z_\ell = T_{d_\ell}(z_{\ell-1}), \qquad z_L = T_{\prod_{\ell=1}^{L} d_\ell}(z_0). \tag{91}$$

No intermediate measurement is required.

Proof. Let z_ℓ denote the centered response produced by layer ℓ. The first layer gives

$$z_1 = T_{d_1}(z_0) \tag{92}$$

by Corollary 10.1. For every $\ell \geq 2$, the one-branch form of Eq. (86) for layer $\ell - 1$ gives

$$\Pr(c_{\ell-1} = 0) = \frac{1 + z_{\ell-1}}{2}. \tag{93}$$

For the next layer, the projector in Eq. (90) therefore has centered expectation $2\Pr(c_{\ell-1} = 0) - 1 = z_{\ell-1}$. Corollary 10.1 yields $z_\ell = T_{d_\ell}(z_{\ell-1})$. Iterating and using $T_a(T_b(t)) = T_{ab}(t)$ proves the second identity. □

The theorem realizes literal Chebyshev composition inside the coherent architecture. It routes one scalar response through the interference registers; the complete multi-branch state may contain additional coherent information.

10.3 A process-space lift for the full deep circuit

A deep current-state reflection depends jointly on the input and all earlier trainable gates. We therefore expand Eq. (83) and unroll the circuit to its actual data-dependent oracle sites. At site a, $\mathcal{X}_x^{(a)}$ denotes the unitary channel physically inserted there. This may be the base channel $\mathcal{U}_x(\cdot) = U_x(\cdot)U_x^\dagger$, its inverse, or a controlled variant if the chosen compilation uses one; coherently controlled access requires a fixed circuit-level phase convention. In the standard reflection decomposition used for the query recurrence below, controlled state reflections use unconditional U_x and $U_x^\dagger$ around a controlled zero-state phase, so no controlled unknown channel is inferred merely from its Choi matrix.

For a channel $\mathcal{X} : L(\mathcal{H}_{\text{in}}) \to L(\mathcal{H}_{\text{out}})$, fix the Choi convention

$$J(\mathcal{X}) = \sum_{r,s} |r\rangle\langle s| \otimes \mathcal{X}(|r\rangle\langle s|). \tag{94}$$

Theorem 10.3 (Unrolled data-oracle tester representation) *Fix a finite QFNN circuit and an output event y. If the unrolled circuit contains r calls to data-dependent channels $\mathcal{X}_x^{(1)}, \ldots, \mathcal{X}_x^{(r)}$, then there is a positive tester operator $\Upsilon_{y,\theta}$, independent of x and satisfying the quantum-comb normalization constraints when summed over y, such that*

$$p_\theta(y \mid x) = \operatorname{Tr}\left[\Upsilon_{y,\theta} \bigotimes_{a=1}^{r} J(\mathcal{X}_x^{(a)})^{\mathsf{T}}\right]. \tag{95}$$

The transpose is taken in the fixed Choi bases; fixed wire permutations are absorbed into the tester.

Proof. Cut the circuit at every base data-oracle call. The remainder is an r-slot quantum network

containing all initial fixed states, trainable gates, controls, memory wires, and the terminal event. The generalized Born rule for quantum combs expresses the outcome probability as the link product of this tester and the Choi matrices inserted in its slots (Chiribella et al., 2009). Under Eq. (94), rewriting the link product as a Hilbert–Schmidt contraction gives Eq. (95). □

Let $n_a = \dim \mathcal{H}_{\mathrm{in},a}$. Because every $\mathcal{X}_x^{(a)}$ in the declared model is a unitary channel, normalize its Choi matrix as $\omega_x^{(a)} = J(\mathcal{X}_x^{(a)})^{\mathsf{T}}/n_a$. The process feature

$$\Omega_x = \bigotimes_{a=1}^{r} \omega_x^{(a)} \tag{96}$$

has Hilbert–Schmidt norm one; for a general noisy channel the corresponding norm is only bounded above by one. All normalization factors can be absorbed into the tester. The associated unrestricted process-feature kernel is

$$\begin{aligned} K_{\mathrm{proc}}(x, x') &= \mathrm{Tr}(\Omega_x \Omega_{x'}) \\ &= \prod_{a=1}^{r} \mathrm{Tr}\left[\omega_x^{(a)} \omega_{x'}^{(a)}\right]. \end{aligned} \tag{97}$$

Corollary 10.4 (Process feature rank and norm complexity) *On a domain X, put*

$$r_{\mathrm{proc}}(X) = \dim_{\mathbb{R}} \mathrm{span}_{\mathbb{R}}\{\Omega_x : x \in X\}. \tag{98}$$

The unrestricted real class $\{x \mapsto \mathrm{Tr}(A\Omega_x) : A = A^\dagger\}$ has linear dimension and pseudo-dimension $r_{\mathrm{proc}}(X)$. On a sample S, its rank equals the rank of the Gram matrix in Eq. (97). Moreover, for $\|A\|_{\mathrm{HS}} \le B$,

$$\widehat{\mathfrak{R}}_S \le \frac{B}{\sqrt{m}}. \tag{99}$$

Proof. The dimension and pseudo-dimension statements repeat the finite-dimensional linear-feature argument of Theorem 6.7. Gram rank equals feature-span rank. Finally,

$$\widehat{\mathfrak{R}}_S = \frac{B}{m} \mathbb{E}_\sigma \left\| \sum_{i=1}^{m} \sigma_i \Omega_{x_i} \right\|_{\mathrm{HS}} \tag{100}$$

$$\le \frac{B}{m} \sqrt{\sum_{i=1}^{m} \|\Omega_{x_i}\|_{\mathrm{HS}}^2} = \frac{B}{\sqrt{m}}. \tag{101}$$

□

This corollary describes the unrestricted Hermitian linear envelope of the process feature. Physical QFNN outcome testers form its positive, normalized, causally constrained subset, so r_{proc} and Eq. (99) give dimension and complexity upper bounds for realizable QFNN outcomes.

If each Choi matrix has a finite Fourier expansion

$$J(\mathcal{X}_x^{(a)}) = \sum_{\lambda \in \Lambda_a} e^{i\lambda \cdot x} C_{a,\lambda}, \tag{102}$$

then

$$\mathrm{supp}\, p_\theta(y \mid \cdot) \subseteq \Lambda_1 + \cdots + \Lambda_r. \tag{103}$$

Thus all attainable output frequencies lie in the stated Minkowski sum; their coefficients are determined by the physical tester.

10.4 What transfers from one layer

The recursive construction retains four conclusions, but through logically distinct arguments and under distinct assumptions.

1. The exact projector-overlap activation law survives at every layer by Corollary 10.1; a generic layer may also retain information beyond its reported scalar responses.
2. A fixed unrolled deep circuit is linear in the enlarged multi-query process feature of Theorem 10.3. Consequently, rank, kernel, Fourier-support, and norm-complexity analyses can be carried out in that process space, subject to the physical tester restrictions stated after Corollary 10.4.
3. The scalar-routing subarchitecture realizes literal Chebyshev composition by Theorem 10.2.
4. The one-layer Universal Approximation Theorem transfers to every prescribed finite coherent depth by the exact-recovery argument below, provided the complete prefix state is retained and the terminal layer permits the reference bank and affine readout stated in Theorem 10.6.

The first three conclusions follow respectively from two-reflection geometry, the process lift, and the routed projector construction. The fourth follows from the exact-recovery lemma and the one-layer Universal Approximation Theorem, as shown next.

10.5 Exact recovery and transfer of approximation

To display the same parameter dependence directly in the response law, write the complete incoming state to layer ℓ as $|\Phi^{(\ell-1)}_{\theta_{<\ell}}(x)\rangle$ and its density operator as $\rho^{(\ell-1)}_{\theta_{<\ell}}(x)$. In the next display, $Q^{(\ell)}_j(\theta_{\ell,j})$ denotes either a global current reference projector or the full-space extension in Eq. (84). Although the arguments of $V^{(\ell)}_j$ are suppressed in Eq. (85), it contains both the data and previous-layer parameters through the current-state reflection, and the current-layer parameters through the reference reflection. The two equivalent response descriptions are

$$\begin{aligned}
&\langle\Phi^{(\ell-1)}_{\theta_{<\ell}}(x)|\, V^{(\ell)}_j\, |\Phi^{(\ell-1)}_{\theta_{<\ell}}(x)\rangle \\
&\qquad = T_{d^{(\ell)}_j}\left(2p^{(\ell)}_j(x)-1\right), \qquad (104)\\
&p^{(\ell)}_j(x) = \mathrm{Tr}\left[\rho^{(\ell-1)}_{\theta_{<\ell}}(x) Q^{(\ell)}_j(\theta_{\ell,j})\right]. \qquad (105)
\end{aligned}$$

For a local reference, the second trace may equivalently be taken against the corresponding reduced state, as in Eq. (87).

The complete layer state contains more information than this reported scalar response. Define the normalized address state and an input-independent linear recovery map by

$$\begin{aligned}
|a_\ell\rangle &= \sum_j \sqrt{w^{(\ell)}_j}\,|j\rangle\,, \\
D_\ell &= \langle a_\ell|_{A_\ell} \otimes (\langle 0| + \langle 1|)_{c_\ell} \otimes \mathbb{I}.
\end{aligned} \qquad (106)$$

Lemma 10.5 (Exact recovery of the complete incoming state) *For every fixed parameter choice and every input x, the full-retention layer state in Eq. (86) satisfies*

$$D_\ell\, |\Phi^{(\ell)}_{\theta_{\le\ell}}(x)\rangle = |\Phi^{(\ell-1)}_{\theta_{<\ell}}(x)\rangle\,. \qquad (107)$$

Equivalently,

$$D_\ell \rho^{(\ell)}_{\theta_{\le\ell}}(x) D^\dagger_\ell = \rho^{(\ell-1)}_{\theta_{<\ell}}(x). \qquad (108)$$

Consequently, if the base encoding $x \mapsto \rho_x$ is injective, then the complete retained-state map is injective after every finite number of layers for every fixed common parameter setting.

Proof. Apply D_ℓ to Eq. (86). Orthogonality of the address words and $\sum_j w^{(\ell)}_j = 1$ give

$$\begin{aligned}
D_\ell\, |\Phi^{(\ell)}_{\theta_{\le\ell}}(x)\rangle &= \sum_j w^{(\ell)}_j \left(\frac{\mathbb{I}+V^{(\ell)}_j}{2} + \frac{\mathbb{I}-V^{(\ell)}_j}{2}\right) \\
&\quad \times |\Phi^{(\ell-1)}_{\theta_{<\ell}}(x)\rangle \\
&= |\Phi^{(\ell-1)}_{\theta_{<\ell}}(x)\rangle\,. \qquad (109)
\end{aligned}$$

Taking outer products proves Eq. (108). Equality of two layer-ℓ density operators therefore implies equality of their layer-$(\ell-1)$ predecessors; induction reduces the claim to injectivity of $x \mapsto \rho_x$. □

The map D_ℓ is an algebraic left inverse, not a deterministic inverse channel. The normalized operator $D_\ell/\sqrt{2}$ is a valid postselection Kraus operator and recovers the incoming state with probability $1/2$. Because Eq. (86) is a finite continuous construction in the incoming pure-state projector for fixed parameters, continuity and purity are also preserved.

Theorem 10.6 (Universal Approximation Theorem at prescribed coherent depth) *Let X be compact Hausdorff and let $x \mapsto \rho_x$ be continuous, injective, and pure-valued. Fix a finite depth $L \ge 1$ and arbitrary fixed parameters for the first $L-1$ full-retention layers. In layer L, allow arbitrary finite width, arbitrary pure reference projectors on the complete retained Hilbert space, arbitrary nonnegative integer reflection degrees, and an unrestricted affine readout. The union of the resulting depth-L models is uniformly dense in $C(X,\mathbb{R})$.*

If X is compact metric and μ is a finite regular Borel measure, the same family is dense in $L^p(X,\mu)$ for every $1 \le p < \infty$. With the matrix head in Eq. (9), the corresponding finite-output family is dense in $C(X,\mathbb{R}^C)$, and in $L^p(X,\mu;\mathbb{R}^C)$ under the latter assumptions.

Proof. By Lemma 10.5, the complete prefix state $x \mapsto \rho^{(L-1)}_{\theta_{<L}}(x)$ is a continuous, injective, pure-valued encoding. Equations (104) and (105) show that the final layer is exactly a one-layer Chebyshev bank on that encoding. Theorem 7.2 and Corollaries 7.4 and 7.5 therefore apply directly. □

Corollary 10.7 (Diagonal-head density after coherent matrix-coefficient compilation) *Under the assumptions*

of Theorem 10.6, fix $L \geq 2$ and arbitrary parameters for the first $L-2$ full-retention layers. In layer $L-1$, allow arbitrary finite class-tagged width, arbitrary pure reference projectors on the complete prefix space, arbitrary reflection degrees, and adjustable address weights. In layer L, allow one degree-one branch per output coordinate whose reference is an orthogonal projector on $A_{L-1}c_{L-1}$, extended by the identity on older registers. Then the union of models with the diagonal terminal calibration

$$z_c(x) = \beta\delta_c(x) + b'_c, \tag{110}$$

where one scale β is shared by all coordinates, is uniformly dense in $C(X, \mathbb{R}^C)$. Under the measure assumptions of Corollary 7.4, it is also dense in $L^p(X, \mu; \mathbb{R}^C)$ for every $1 \leq p < \infty$.

Proof. Apply scalar density coordinatewise on the continuous injective pure-state encoding produced by the fixed prefix. This gives finite banks

$$h_c(x) = b_c + \sum_{r=1}^{R_c} a_{c,r}\alpha_{c,r}(x). \tag{111}$$

Discard all zero-coefficient atoms from the banks. Put $S_c = \sum_r |a_{c,r}|$ and $S = \sum_c S_c$. For $S > 0$, concatenate the banks into one layer indexed by (c, r) and choose its address probabilities

$$w_{c,r} = \frac{|a_{c,r}|}{S}. \tag{112}$$

For every remaining coefficient choose $\tau_{c,r} \in \{0, 1\}$ with $(-1)^{\tau_{c,r}} = \operatorname{sgn}(a_{c,r})$, and define the higher-rank local projector

$$P_c = \sum_r \big(|c, r\rangle\langle c, r|\big)_{A_{L-1}} \otimes \big(|\tau_{c,r}\rangle\langle\tau_{c,r}|\big)_{c_{L-1}} \otimes \mathbb{I}_{\text{old}}. \tag{113}$$

The P_c are mutually orthogonal. The retained-state amplitudes in Eq. (86) give the joint probabilities

$$\begin{aligned} &\Pr(A_{L-1} = (c, r),\, c_{L-1} = \tau) \\ &\quad = \frac{w_{c,r}}{2}\left[1 + (-1)^{\tau}\alpha_{c,r}(x)\right]. \end{aligned} \tag{114}$$

Summing the interference arm selected by $\tau_{c,r}$ therefore yields

$$\begin{aligned} p_c(x) &= \operatorname{Tr}\left[\rho^{(L-1)}(x)P_c\right] \\ &= \frac{S_c}{2S} + \frac{1}{2S}\sum_r a_{c,r}\alpha_{c,r}(x). \end{aligned} \tag{115}$$

The degree-one response is therefore

$$\delta_c(x) = 2p_c(x) - 1 = -1 + \frac{S_c}{S} + \frac{1}{S}\sum_r a_{c,r}\alpha_{c,r}(x). \tag{116}$$

Choosing $\beta = S$ and $b'_c = b_c + S - S_c$ makes $\beta\delta_c + b'_c = h_c$ exactly. If $S_c = 0$, take $P_c = 0$ and the same bias formula; here the zero projector is allowed, its reflection is the identity, and its degree-one centered response is -1. If $S = 0$, all coordinates are constant and $\beta = 0$ suffices. Uniform and L^p density now follow from the componentwise approximants. □

Corollary 10.7 implements the coefficients of a finite multiclass affine bank coherently through address amplitudes and higher-rank terminal projectors. It replaces the full matrix readout by one shared scale and class biases, using class-tagged features and one additional degree-one layer. The coefficient information is thereby moved into address preparation and subspace reflections; the uniform-address numerical ansatz in Appendix B uses the smaller direct calibration instead.

11 Depth and resource accounting

Depth is attractive for the same structural reason as in classical neural networks: compositions can represent high algebraic degree from low-degree local maps. Theorem 10.2 makes that statement exact for a routed QFNN chain. Whether it is an efficiency advantage depends on what is counted.

11.1 A conditional depth–degree separation in a reflection-oracle model

The next elementary lemma supplies a sharp shallow lower bound for the scalar target used in the routed chain.

Lemma 11.1 (Sharp lower-degree approximation barrier) *For every $D \geq 1$,*

$$\inf_{\substack{q \in \mathbb{R}[t] \\ \deg q < D}} \|T_D - q\|_{\infty,[-1,1]} = 1. \tag{117}$$

Proof. The choice $q = 0$ gives the upper bound one. For the reverse inequality, consider $x_k =$

$\cos(k\pi/D)$, $k = 0, \dots, D$, at which $T_D(x_k) = (-1)^k$. If $\|T_D - q\|_\infty < 1$, then $q(x_k)$ has sign $(-1)^k$. It therefore has a zero in every open interval between consecutive x_k, hence at least D distinct zeros. This contradicts $\deg q < D$. □

Proposition 11.2 (Width-robust extracted-response degree barrier and conditional routed separation) *For every integer $D \geq 1$, let $\mathcal{H}$ be finite-dimensional with $\dim \mathcal{H} \geq 2$, fix a pure reference $Q = |\chi\rangle\langle\chi|$, and define on the full pure-state domain*

$$g_D(\rho_\psi) = T_D\Big(2|\langle\chi|\psi\rangle|^2 - 1\Big). \tag{118}$$

Then

$$\inf_{f \in \mathcal{V}_{D-1}} \sup_{\|\psi\|=1} |g_D(\rho_\psi) - f(\rho_\psi)| = 1. \tag{119}$$

Consequently, no extracted-response affine-head bank of the form Eq. (57) whose branches all have degree below D—regardless of its finite width or its reference directions—achieves uniform error below one.

Now work in a high-level oracle model in which reflection about the currently prepared state costs one unit, independently of its preparation cost. A routed L-layer chain with degrees $d_1, \dots, d_L \geq 2$ and $D = \prod_\ell d_\ell$ realizes g_D using $\sum_\ell d_\ell$ alternating reflection pairs. Any exact one-layer extracted-response affine-head bank must contain a branch of degree at least D, and therefore use at least D pairs on that branch. For $d_\ell = d$, the high-level path counts are Ld versus at least d^L.

Proof. The zero function gives the upper bound one. For the lower bound, let $K = D - 1$, take $f \in \mathcal{V}_K$, and Haar-average it over the stabilizer

$$\mathcal{G}_\chi = \{U \in \mathrm{U}(\mathcal{H}) : U|\chi\rangle = |\chi\rangle\}. \tag{120}$$

Define $\overline{f}(\rho_\psi) = \int_{\mathcal{G}_\chi} f(\rho_{U\psi})\, dU$. The target g_D is invariant under this group, so averaging cannot increase the uniform error. It also leaves $\mathcal{V}_K$ invariant, because it merely twirls the Hermitian coefficient operator on $\mathrm{Sym}^K \mathcal{H}$.

Under $\mathcal{H} = \mathbb{C}|\chi\rangle \oplus \chi^\perp$, decompose

$$\mathrm{Sym}^K \mathcal{H} = \bigoplus_{r=0}^{K} |\chi\rangle^{\odot r} \odot \mathrm{Sym}^{K-r}(\chi^\perp). \tag{121}$$

For $K = 0$ this is the single constant sector. In general these sectors are pairwise inequivalent irreducible $\mathrm{U}(\chi^\perp)$ modules: their central characters distinguish them, including the one-dimensional $\chi^\perp$ case. Schur's lemma therefore makes the stabilizer twirl scalar on each sector. Its expectation on $|\psi\rangle^{\otimes K}$ is a linear combination of

$$\binom{K}{r} t^r (1-t)^{K-r}, \qquad 0 \leq r \leq K, \qquad t = |\langle\chi|\psi\rangle|^2. \tag{122}$$

Thus $\overline{f}(\rho_\psi) = q(t)$ for a polynomial q of degree at most K. Every $t \in [0,1]$ occurs when $\dim \mathcal{H} \geq 2$. After the affine change $s = 2t - 1$, Lemma 11.1 gives uniform error at least one, proving Eq. (119). The constructive routed count follows from Theorem 10.2. Finally, every extracted-response affine-head bank of maximum degree below D belongs to $\mathcal{V}_{D-1}$ by Theorem 6.2; hence an exact bank in this class has maximum degree at least D. □

The lower bound holds on the complete space $\mathcal{V}_{D-1}$ and therefore simultaneously covers every finite width and every choice of reference directions. Indeed, by Theorem 6.2, any bank of extracted responses of maximum degree below D, followed by an arbitrary affine readout, is an element of $\mathcal{V}_{D-1}$; the bias and feature coefficients are absorbed into its Hermitian operator. The routed deep construction measures g_D directly. Comparator classes with nonlinear classical post-processing or arbitrary joint measurements of the complete hidden state require a separate analysis.

11.2 Base-oracle accounting

Proposition 11.3 (Preparation–unpreparation query recurrence) *Assume layer ℓ has common coherent path degree d_ℓ, and implement each current-state reflection through Eq. (83). Let q_ℓ count calls to U_x or $U_x^\dagger$ in one execution of the state preparation through layer ℓ, with $q_0 = 1$. Then*

$$q_\ell = (1 + 2d_\ell) q_{\ell-1}, \qquad q_L = \prod_{\ell=1}^{L} (1 + 2d_\ell). \tag{123}$$

Proof. Preparing the incoming state once costs $q_{\ell-1}$ calls. Every one of the d_ℓ appearances of its reflection invokes the previous preparation and inverse, costing $2q_{\ell-1}$. Reference reflections add gates but no base data-oracle calls. □

This count assumes no cached additional copies, cancellation across reflections, or alternative native

reflection primitive. A controlled base-oracle call, if used by a different compilation, must be counted and must obey the fixed-phase access convention stated in Section 10.3.

For scalar routing, a direct shallow degree-D unit costs $1 + 2D$ base calls under the same implementation, whereas the recursively compiled deep chain costs $\prod_\ell (1 + 2d_\ell)$, which is larger for $L \geq 2$ and $d_\ell \geq 1$. Thus the conditional saving of Proposition 11.2 requires a hardware-native current-state reflection, cached coherent resource, approximate reflection, or another compilation model whose cost is separately justified. In standard quantum signal-processing and singular-value-transformation constructions, polynomial degree scales with the number of signal-oracle uses; Eq. (123) is the corresponding architecture-specific accounting for exact current-state reflections (Low and Chuang, 2017; Gilyén et al., 2019). The broader polynomial method likewise relates quantum query complexity to approximation degree (Beals et al., 2001).

11.3 Qubits, width, and implementation resources

The routed construction stores one interference control per layer if all hidden states are retained literally. It therefore uses approximately $q_{\text{work}} + L$ logical qubits, while a direct single-unit degree-D implementation uses $q_{\text{work}} + 1$. The explicit construction therefore trades additional retained controls for coherent recursion. Control recycling would require a different coherent recompilation or a measurement/reset strategy and hence a corresponding resource analysis.

An address register labels R branches with $\lceil \log_2 R \rceil$ qubits, while a generic multiplexed block still requires the descriptions or controlled implementations of R reference preparations. Estimating all branch responses also contributes shot cost. Thus width, control synthesis, memory, and readout accompany the logarithmic address register.

Proposition 11.2 supplies an arbitrary-width lower-degree barrier for one routed target. A full multi-branch depth–width theorem under implementation-level resources would add an explicit target family, a constructive deep circuit, and a lower bound for every shallower competitor under a common measure of reference descriptions, oracle calls, gates, ancillas, shots, and terminal post-processing. The classical results provide useful models for formulating such a theorem (Mhaskar and Poggio, 2016; Telgarsky, 2016; Eldan and Shamir, 2016; Daniely, 2017).

11.4 Trainability and noise

Bounded activations do not imply bounded parameter sensitivity. For $d \geq 1$, writing $z = 2F - 1$ gives

$$\frac{\partial}{\partial \theta} T_d(z) = 2d\, U_{d-1}(z) \frac{\partial F}{\partial \theta}, \tag{124}$$

where U_{d-1} is the Chebyshev polynomial of the second kind. For $d \geq 2$ the factor has $d-1$ interior zeros, and for $d \geq 1$ its magnitude can scale as d^2 near $z = \pm 1$. It can therefore attenuate or amplify an existing overlap gradient depending on the data, degree, and parameters.

Local cost observables can avoid exponential concentration for some shallow, local ansatzes whose causal cones remain small (Cerezo et al., 2021). Although a layer projector may be local, its expectation is taken after a current-state reflection that can be global and recursively deep. Terminal subsystem measurement alone is therefore insufficient to invoke those favourable results. Random-circuit concentration remains a risk (McClean et al., 2018). Under the local Pauli-noise assumptions of Wang et al. (2021), noise can also induce exponentially vanishing gradients with circuit depth. This conclusion is noise-model dependent: for typical circuits with local non-unital noise and local observables, recent work proves absence of barren plateaus while also showing effective logarithmic depth and corresponding classical-simulation consequences (Mele et al., 2026). Equation (124) therefore makes trainability depend jointly on the overlap geometry, reflection degree, data distribution, and circuit ansatz. Because the recursively implemented current-state reflection can be global, analysing gradient concentration and noise requires the causal structure of the complete architecture, even when the measured projector is local.

12 Relation to enlarged kernel models

The one-layer and deep representations fit naturally into the enlarged-linear view of quantum learning models (Schuld, 2021; Jerbi et al., 2023). For one

layer, the appropriate feature is $\rho_x^{\otimes D}$:

$$f(x) = \mathrm{Tr}(A\rho_x^{\otimes D}), \tag{125}$$

with polynomial-fidelity kernel $[\mathrm{Tr}(\rho_x \rho_{x'})]^D$. For a deep circuit, unrolling the data-oracle calls produces the multi-query process feature Ω_x of Theorem 10.3. These two lifts identify the linear spaces in which the respective QFNN models operate while preserving the architecture's explicit projector-overlap response and coherent recursive realization. Their operational cost depends on how the lifted feature or kernel is evaluated.

Compared with data re-uploading, the QFNN provides a neuron-level response $T_d(2p-1)$ and an exact finite-degree capacity theory. Data re-uploading offers a more general circuit template, whereas the present construction provides a more explicit activation and layerwise projector geometry. Their relative performance is therefore a natural subject for matched empirical comparison.

13 Discussion and outlook

The results expose two complementary design axes. Reflection degree enlarges the exact one-layer polynomial space and can reduce target-specific approximation error, while norm and margin control determine the associated statistical complexity. Coherent depth preserves the activation and approximation framework and introduces exact compositional mechanisms whose implementation cost depends on the available reflection primitives.

The next empirical step is a matched study of width, degree, and depth under common data access, gate, qubit, shot, and optimization budgets. The exact tensor and process representations developed here provide the feature ranks, kernels, margins, and oracle counts needed for such comparisons, including future studies of gradient scaling and realistic noise.

14 Conclusion

We introduced a coherent quantum feedforward architecture built from a simple projector-overlap primitive. A trainable projector defines an overlap pre-activation, alternating reflections transform it exactly into the Chebyshev response $T_d(2p-1)$, and an addressed bank of such responses forms a quantum analogue of a one-hidden-layer activation bank.

For one layer, the resulting theory is exact. Degree-at-most-D models span the Hermitian linear functionals of $\rho^{\otimes D}$ on the symmetric subspace. This characterization yields the polynomial-fidelity kernel, strict degree hierarchy, domain-restricted rank and pseudo-dimension, interpolation and finite-separation results, and Fourier structure. It also supports target-specific approximation bounds, Universal Approximation Theorems for scalar and finite-dimensional outputs, and norm- and margin-sensitive generalization bounds.

The same construction extends coherently to depth. Every recursive layer retains the exact projector-overlap activation law, and complete state retention transfers the idealized Universal Approximation Theorem to any prescribed finite depth. The process lift identifies the corresponding multi-query feature representation, while the routed construction demonstrates exact multiplication of Chebyshev degrees and a sharp lower-degree shallow barrier. The simulator experiments on MNIST8, Fashion-MNIST8, and KMNIST8 confirm that the proposed architecture can be trained in a multiclass setting. Together, these results establish a concrete mathematical bridge between neural-style polynomial activations and coherent variational quantum circuits, and provide a foundation for future resource-matched studies of depth, trainability, and noise.

A Circuit schematics

The diagrams in this appendix expose the register structure suppressed in the algebraic definition. They are oracle-level schematics, not elementary-gate decompositions. A filled control on an interference qubit c_ℓ means that the corresponding block is applied on the $|1\rangle_{c_\ell}$ arm and the identity is applied on the $|0\rangle_{c_\ell}$ arm. We write

$$\mathcal{C}_c[U] = |0\rangle\langle 0|_c \otimes \mathbb{I} + |1\rangle\langle 1|_c \otimes U \tag{126}$$

for this control. The address-selected reference reflection is the block-diagonal multiplexor

$$\mathrm{SELECT}_A(R_Q) = \sum_{j=0}^{2^k-1} |j\rangle\langle j|_A \otimes \widetilde{R}_j,$$
$$\widetilde{R}_j = \begin{cases} R_{Q_j} = \mathbb{I} - 2Q_j, & 0 \le j < R, \\ \mathbb{I}, & R \le j < 2^k. \end{cases} \tag{127}$$

Thus address word $j < R$ selects exactly one reflection R_{Q_j}; the reflections are not applied sequentially. Identity blocks complete the unitary on unused address words when $R < 2^k$ and are absent when $R = 2^k$. The address-control symbol in the figures denotes this whole-register multiplexor, not a control only on the all-ones address word.

The reported numerical model has $R = 2^k = 8$ and uses $U_A(w) = H^{\otimes k}$, hence $w_j = 1/8$. A more general $U_A(w)$ permits nonuniform weights. Conditional branch responses are unchanged after division by w_j, although off-diagonal address coherences can matter when the register is retained for a later layer. The schematic assumes a common branch degree d. Unequal degrees require an address-controlled schedule that activates each branch for its prescribed number of rounds.

For $w_j > 0$, marginalizing the work register gives

$$\Pr(A = j, c = b) = w_j \frac{1 + (-1)^b \alpha_j}{2}, \tag{128}$$
$$\alpha_j = T_d(2\,\langle\phi_x|\,Q_j\,|\phi_x\rangle - 1)\,. \tag{129}$$

Consequently,

$$\alpha_j = \frac{\Pr(j,0) - \Pr(j,1)}{\Pr(j,0) + \Pr(j,1)}. \tag{130}$$

Measuring only c returns the weighted aggregate $\sum_j w_j \alpha_j$. Recovering the individual response bank requires the joint (A, c) readout, namely $k + 1$ measured qubits, but not measurement of W.

For the recursive diagram, let the complete retained state after the first layer be $|\Phi_x^{(1)}\rangle$ on

$$\mathcal{R}_1 = W \otimes S_1, \qquad S_1 = A_1 \otimes c_1, \tag{131}$$

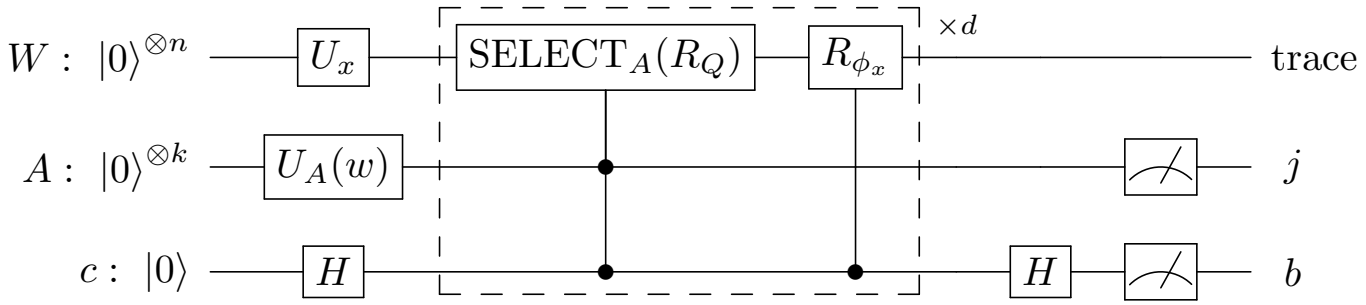
and let $U_1(x, \theta^{(1)})$ denote the complete first-layer preparation, including both the data encoding and the trainable first-layer references: $U_1(x, \theta^{(1)})\,|0\rangle = |\Phi_x^{(1)}\rangle$. A second-layer projector $Q_j^{(2)}$ on S_1 is extended to $\widetilde{Q}_j^{(2)} = \mathbb{I}_W \otimes Q_j^{(2)}$ on $\mathcal{R}_1$.

At the terminal layer,

$$\Pr(A_2 = j, c_2 = b) = w_j^{(2)} \frac{1 + (-1)^b \alpha_j^{(2)}}{2}, \tag{132}$$
$$\alpha_j^{(2)} = T_{d_2}\left(2p_j^{(2)} - 1\right), \tag{133}$$
$$p_j^{(2)} = \langle\Phi_x^{(1)}|\,\widetilde{Q}_j^{(2)}\,|\Phi_x^{(1)}\rangle\,. \tag{134}$$

Thus, for every terminal branch with $w_j^{(2)} > 0$, $k_2 + 1$ measured qubits suffice. Measuring the retained registers as well and summing their counts gives the same marginal but is optional. Keeping those joint outcomes as additional features would define a richer readout than the response bank analyzed here.

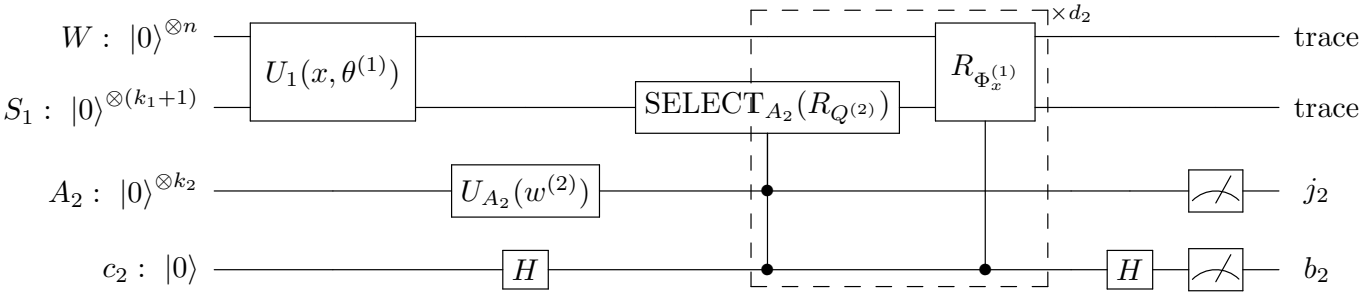
The reported experiment is unphased. Appendix C records a short conceptual phase-modulation extension; it is not part of the implemented model.



$$U_x\,|0\rangle^{\otimes n} = |\phi_x\rangle\,, \qquad U_A(w)\,|0\rangle^{\otimes k} = \sum_j \sqrt{w_j}\,|j\rangle\,,$$

$$R_{\phi_x} = \mathbb{I} - 2\,|\phi_x\rangle\langle\phi_x|\,, \qquad V_j^{(d)} = (R_{\phi_x} R_{Q_j})^d.$$

Fig. 1 One addressed QFNN layer with common reflection degree d. The dashed box is one reflection round and is repeated d times, so address branch j implements $(R_{\phi_x} R_{Q_j})^d$. Both reflections are enabled by $c = 1$, while A selects the reference reflection. The meters show terminal readout of the address and interference registers; the work register is marginalized. When this block is used as a hidden layer, the terminal meters are omitted and the work, address, and interference registers are passed coherently to the next layer.



$$|\Phi_x^{(1)}\rangle = U_1(x,\theta^{(1)})\,|0\rangle_{\mathcal{R}_1}\,, \qquad \widetilde{Q}_j^{(2)} = \mathbb{I}_W \otimes Q_j^{(2)},$$

$$R_{\widetilde{Q}_j^{(2)}} = \mathbb{I}_{\mathcal{R}_1} - 2\widetilde{Q}_j^{(2)}, \qquad R_{\Phi_x^{(1)}} = \mathbb{I}_{\mathcal{R}_1} - 2\,|\Phi_x^{(1)}\rangle\langle\Phi_x^{(1)}|\,.$$

Fig. 2 A recursive second layer. The first block prepares the complete coherent first-layer state. In each of the d_2 rounds enclosed by the dashed box, the selected reference reflection acts locally on $S_1 = A_1 c_1$ and as the identity on W, while the following current-state reflection acts globally on $W A_1 c_1$. Both are enabled by $c_2 = 1$, and A_2 selects the reference reflection. The terminal joint measurement of (A_2, c_2) extracts the second-layer response bank while older registers are marginalized. For a further hidden layer, the meters are omitted and all displayed registers remain coherent.

B Implemented numerical experiment protocol

MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), and KMNIST (Clanuwat et al., 2018) are restricted to labels $0, \ldots, 7$. Images are converted to grayscale, resized to 8×8, flattened, and Euclidean-normalized to define a real six-qubit amplitude state. For each dataset, a stratified split reserves 200 examples per class (1600 in total) from the official training partition for validation. The final train/validation/test counts are 8000/1600/8017 for MNIST8 and 8000/1600/8000 for each of Fashion-MNIST8 and KMNIST8.

Model and objective. Each layer uses eight trainable real reference states, a uniform three-qubit address register, one interference qubit, and a common degree $d \in \{1, 2, 3\}$. Each reference uses an initial R_y layer followed by four CNOT-ring/R_y repetitions. In the two-layer model the reference reflection acts locally on $A_1 c_1$, the current-state reflection acts globally on the retained $W A_1 c_1$, and no intermediate measurement or re-encoding occurs. The one- and two-layer models contain respectively 249 and 409 trainable parameters, including the terminal calibration

$$z_j = \beta \alpha_j + b_j, \qquad \alpha_j = 8\,[p(j,0) - p(j,1)]. \tag{135}$$

The implemented architecture is unphased and contains no trainable interference-phase parameter. For label smoothing ε and entropy weight λ_H, the training objective is

$$q_j = (1-\varepsilon)\mathbf{1}_{j=y} + \frac{\varepsilon}{8}, \tag{136}$$

$$\pi = \mathrm{softmax}(z), \tag{137}$$

$$\mathcal{H}(\pi) = -\sum_{j=0}^{7} \pi_j \log \pi_j, \tag{138}$$

$$\mathcal{L} = -\sum_{j=0}^{7} q_j \log \pi_j + \lambda_H \mathcal{H}(\pi). \tag{139}$$

A positive λ_H therefore penalizes predictive entropy.

For each input and forward pass, training uses one multinomial draw with $N_{\mathrm{shot}} = 6000$ from the complete terminal address–interference distribution. The sampled response is used in the forward loss, while a straight-through estimator based on the exact-statevector response supplies the gradient. Validation uses exact probabilities.

Search and final evaluation. For each dataset, we ran an independent 60-trial Optuna (Akiba et al., 2019) study with the tree-structured Parzen estimator (Bergstra et al., 2011). Each trial trained two seeds, 101 and 102, for 30 epochs and maximized their mean best exact-validation accuracy. The search included

$$(L, d) \in \{1, 2\} \times \{1, 2, 3\},$$

optimizer, learning-rate schedule, batch size, weight decay, label smoothing, and entropy regularization. The first six trials covered all depth–degree pairs once, after which the remaining trials were selected by the sampler. Search runs used at most 300 training examples per class.

The selected recipe for each dataset was subsequently trained for 100 epochs with seeds $41, \ldots, 46$, using at most 1000 training examples per class. Each seed's validation-selected checkpoint was evaluated on the official test partition using exact probabilities and five repeated 6000-shot evaluations. The five finite-shot accuracies were first averaged within each training seed; reported means and sample standard deviations were then computed across the six independently trained seeds. The plotted trajectories contain online finite-shot training accuracy and exact validation accuracy.

Reproducibility. The repository cited in the Code availability statement contains the complete implementation and experimental record, including the fixed run configuration, optimization studies, selected recipes, per-seed histories and checkpoints, aggregate result tables, and plotting data.

B.1 Proof-of-concept results

All three validation-selected recipes use two coherent layers and 409 trainable parameters. MNIST8 selects degree three, whereas Fashion-MNIST8 and KMNIST8 select degree two. Their exact test accuracies are $90.57 \pm 0.20\%$, $74.55 \pm 0.51\%$, and $67.61 \pm 0.65\%$, respectively. Replacing exact terminal probabilities by 6000-shot multinomial estimates changes the corresponding means by 0.06, 0.59, and 0.40 percentage points. The trajectories in Fig. 3 show stable optimization across the six training seeds. KMNIST8 exhibits a larger validation–test gap than the other two datasets, motivating evaluation on additional splits in future benchmark studies.

Table 1 Proof-of-concept eight-class results. All accuracies are percentages and are reported as mean ± sample standard deviation over six independently trained seeds. For finite-shot test accuracy, five 6000-shot evaluations are averaged within each seed before aggregation across seeds.

Dataset	(L, d)	Online train at selected epoch	Best exact validation	Exact test	6000-shot test
MNIST8	$(2, 3)$	90.64 ± 0.36	89.50 ± 0.59	90.57 ± 0.20	90.51 ± 0.17
Fashion-MNIST8	$(2, 2)$	75.75 ± 0.53	77.17 ± 0.72	74.55 ± 0.51	73.96 ± 0.48
KMNIST8	$(2, 2)$	80.50 ± 0.29	82.19 ± 0.59	67.61 ± 0.65	67.21 ± 0.69

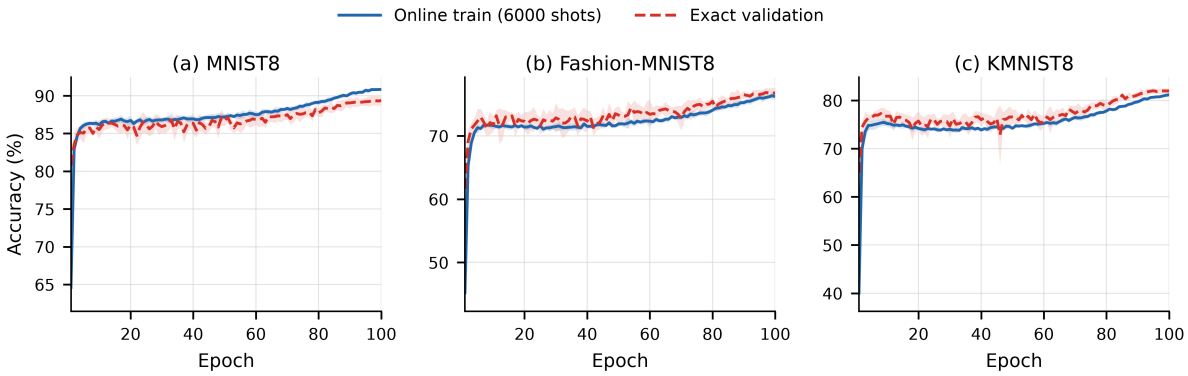


Fig. 3 Training trajectories for the validation-selected QFNN recipes, shown as mean ± sample standard deviation over six independently trained seeds. Solid blue curves denote online training accuracy obtained with 6000-shot multinomial sampling; dashed red curves denote exact validation accuracy.

C Interference-phase modulation

A single $R_z(\varphi)$ on the shared interference qubit before the final Hadamard attaches the same coherent coefficient to every addressed branch response. Up to a global phase, the branch-one arm acquires $e^{i\varphi}$, so

$$\Pr(c=0) = \frac{1 + \mathrm{Re}[e^{i\varphi} \langle\psi| V |\psi\rangle]}{2}. \tag{140}$$

For the unphased reflection sequence, $\langle\psi| V |\psi\rangle = T_d(2p-1)$ is real, and the centered response becomes

$$\cos(\varphi) T_d(2p-1). \tag{141}$$

This phase modulation provides a simple coherent coefficient for a branch response. It is particularly useful between deep QFNN layers, where such coefficients can be introduced without measuring and re-encoding the hidden state. Independent branch coefficients require the address-controlled operation

$$\sum_j |j\rangle\langle j|_A \otimes R_z(\varphi_j), \tag{142}$$

which yields $\cos(\varphi_j) T_d(2p_j - 1)$ in branch j. This modulation preserves the polynomial degree. The Universal Approximation Theorems use the more general affine combinations stated in their hypotheses. The numerical experiment is unphased.

D Numerical sanity checks

Independent small-dimensional calculations were used as diagnostics, not as substitutes for the proofs. Random finite-dimensional projectors verified Theorem 4.1 to floating-point precision for multiple ranks and degrees. Sampled coherent projectors achieved ranks 9 for $(n, D) = (2, 2)$, 16 for $(2, 3)$, and 36 for $(3, 2)$, matching $\binom{n+D-1}{D}^2$. The interpolation construction reproduced Kronecker values on random distinct pure states, and exhaustive linear-separator checks for the equatorial example gave the predicted $D-1$ minimum errors. Random pure-state Gram matrices also satisfied the nondecreasing effective-dimension property in Proposition 9.4 across the tested degrees and ridge values. Random finite multiclass banks satisfied the diagonal-head compilation identity in Corollary 10.7 with maximum discrepancy 7.2×10^{-15}.

The accompanying sanity-check script reproduces these diagnostics and is available in the repository cited below; the analytic theorem statements are established by the proofs in the main text.

Statements and Declarations

Funding

No funds, grants, or other support were received.

Competing interests

The author has no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Clinical trial number

Not applicable.

Data availability

No new benchmark dataset was created for this study. The numerical experiments use the publicly available MNIST, Fashion-MNIST, and KMNIST datasets, restricted to their original labels $0, \ldots, 7$. The research data generated during this study—including run configurations, hyperparameter-optimization records, selected recipes, per-seed histories and checkpoints, aggregate result tables, and plotting data—are available in the public repository https://github.com/KiyoLelou10/QFNN.

Code availability

The QFNN implementation, theorem sanity checks, and complete reproducibility materials for the numerical experiments are available in the public repository https://github.com/KiyoLelou10/QFNN. The archived materials include run configurations, optimization records, selected recipes,

per-seed histories and checkpoints, aggregate result tables, and plotting data.

Author contributions

The author conceived the study, developed the architecture and theoretical results, implemented and validated the software, conducted and analyzed the numerical experiments, and wrote and revised the manuscript.

Use of generative AI

The author used OpenAI ChatGPT and Codex as assistive tools for manuscript organization and drafting, literature searches, code development, mathematical cross-checking, and language editing. The author critically reviewed and verified the final manuscript, references, proofs, software, and numerical results and takes full responsibility for the content.

References

Abbas A, Sutter D, Zoufal C, Lucchi A, Figalli A, Woerner S (2021) The power of quantum neural networks. Nat Comput Sci 1:403–409. https://doi.org/10.1038/s43588-021-00084-1

Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2623–2631. https://doi.org/10.1145/3292500.3330701

Bartlett PL, Mendelson S (2002) Rademacher and Gaussian complexities: risk bounds and structural results. J Mach Learn Res 3:463–482. https://www.jmlr.org/papers/v3/bartlett02a.html

Bergstra JS, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems 24. https://papers.nips.cc/paper/4443

Beals R, Buhrman H, Cleve R, Mosca M, de Wolf R (2001) Quantum lower bounds by polynomials. J ACM 48:778–797. https://doi.org/10.1145/502090.502097

Beer K, Bondarenko D, Farrelly T, Osborne TJ, Salzmann R, Scheiermann D, Wolf R (2020) Training deep quantum neural networks. Nat Commun 11:808. https://doi.org/10.1038/s41467-020-14454-2

Brassard G, Høyer P, Mosca M, Tapp A (2002) Quantum amplitude amplification and estimation. In: Lomonaco SJ, Brandt HE (eds) Quantum Computation and Information. AMS, Providence, pp 53–74. https://doi.org/10.1090/conm/305/05215

Cao Y, Guerreschi GG, Aspuru-Guzik A (2017) Quantum neuron: an elementary building block for machine learning on quantum computers. arXiv:1711.11240. https://doi.org/10.48550/arXiv.1711.11240

Caro MC, Huang HY, Cerezo M, et al (2022) Generalization in quantum machine learning from few training data. Nat Commun 13:4919. https://doi.org/10.1038/s41467-022-32550-3

Cerezo M, Sone A, Volkoff T, Cincio L, Coles PJ (2021) Cost function dependent barren plateaus in shallow parametrized quantum circuits. Nat Commun 12:1791. https://doi.org/10.1038/s41467-021-21728-w

Chiribella G, D'Ariano GM, Perinotti P (2009) Theoretical framework for quantum networks. Phys Rev A 80:022339. https://doi.org/10.1103/PhysRevA.80.022339

Clanuwat T, Bober-Irizar M, Kitamoto A, Lamb A, Yamamoto K, Ha D (2018) Deep learning for classical Japanese literature. arXiv:1812.01718. https://doi.org/10.48550/arXiv.1812.01718

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Math Control Signals Syst 2:303–314. https://doi.org/10.1007/BF02551274

Daniely A (2017) Depth separation for neural networks. In: Proceedings of the 30th Conference on Learning Theory, PMLR 65:690–696. https://proceedings.mlr.press/v65/daniely17a.html

Eldan R, Shamir O (2016) The power of depth for feedforward neural networks. In: Proceedings of

the 29th Annual Conference on Learning Theory, PMLR 49:907–940. https://proceedings.mlr.press/v49/eldan16.html

Gil-Fuster E, Eisert J, Bravo-Prieto C (2024) Understanding quantum machine learning also requires rethinking generalization. Nat Commun 15:2277. https://doi.org/10.1038/s41467-024-45882-z

Gilyén A, Su Y, Low GH, Wiebe N (2019) Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In: Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, pp 193–204. https://doi.org/10.1145/3313276.3316366

Hur T, Park DK (2025) Understanding generalization in quantum machine learning with margins. In: Proceedings of the 42nd International Conference on Machine Learning, PMLR 267:26338–26360. https://proceedings.mlr.press/v267/hur25a.html

Iten R, Colbeck R, Kukuljan I, Home J, Christandl M (2016) Quantum circuits for isometries. Phys Rev A 93:032318. https://doi.org/10.1103/PhysRevA.93.032318

Ivashkov P, Huang PW, Koor K, Pira L, Rebentrost P (2026) QKAN: quantum Kolmogorov–Arnold networks with applications in machine learning and multivariate state preparation. npj Quantum Inf 12:73. https://doi.org/10.1038/s41534-026-01202-5

Jerbi S, Fiderer LJ, Poulsen Nautrup H, Kübler JM, Briegel HJ, Dunjko V (2023) Quantum machine learning beyond kernel methods. Nat Commun 14:517. https://doi.org/10.1038/s41467-023-36159-y

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791

Li S, Salek MS, Wang Y, Chowdhury M (2024) Quantum-inspired activation functions and quantum Chebyshev-polynomial network. arXiv:2404.05901. https://doi.org/10.48550/arXiv.2404.05901

Low GH, Chuang IL (2017) Optimal Hamiltonian simulation by quantum signal processing. Phys Rev Lett 118:010501. https://doi.org/10.1103/PhysRevLett.118.010501

Maronese M, Destri C, Prati E (2022) Quantum activation functions for quantum neural networks. Quantum Inf Process 21:128. https://doi.org/10.1007/s11128-022-03466-0

McClean JR, Boixo S, Smelyanskiy VN, Babbush R, Neven H (2018) Barren plateaus in quantum neural network training landscapes. Nat Commun 9:4812. https://doi.org/10.1038/s41467-018-07090-4

Mele AA, Angrisani A, Ghosh S, Khatri S, Eisert J, Stilck França D, Quek Y (2026) Noise-induced shallow circuits and the absence of barren plateaus. Nat Phys 22:751–756. https://doi.org/10.1038/s41567-026-03245-z

Mhaskar H, Poggio T (2016) Deep vs. shallow networks: an approximation theory perspective. Anal Appl 14:829–848. https://doi.org/10.1142/S0219530516400042

Pérez-Salinas A, Cervera-Lierta A, Gil-Fuster E, Latorre JI (2020) Data re-uploading for a universal quantum classifier. Quantum 4:226. https://doi.org/10.22331/q-2020-02-06-226

Schuld M, Killoran N (2019) Quantum machine learning in feature Hilbert spaces. Phys Rev Lett 122:040504. https://doi.org/10.1103/PhysRevLett.122.040504

Schuld M (2021) Supervised quantum machine learning models are kernel methods. arXiv:2101.11020. https://doi.org/10.48550/arXiv.2101.11020

Schuld M, Sweke R, Meyer JJ (2021) Effect of data encoding on the expressive power of variational quantum-machine-learning models. Phys Rev A 103:032430. https://doi.org/10.1103/PhysRevA.103.032430

Singh U, Goldberg AZ, Heshami K (2024) Coherent feed-forward quantum neural network. Quantum Mach Intell 6:89. https://doi.org/10.1007/s42484-024-00222-8

Stone MH (1948) The generalized Weierstrass approximation theorem. Math Mag 21:167–184. https://doi.org/10.2307/3029750

Tacchino F, Barkoutsos P, Macchiavello C, Tavernelli I, Gerace D, Bajoni D (2020) Quantum implementation of an artificial feed-forward neural network. Quantum Sci Technol 5:044010. https://doi.org/10.1088/2058-9565/abb8e4

Telgarsky M (2016) Benefits of depth in neural networks. In: Proceedings of the 29th Annual Conference on Learning Theory, PMLR 49:1517–1539. https://proceedings.mlr.press/v49/telgarsky16.html

Trefethen LN (2019) Approximation Theory and Approximation Practice, extended edn. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611975949

Wang S, Fontana E, Cerezo M, et al (2021) Noise-induced barren plateaus in variational quantum algorithms. Nat Commun 12:6961. https://doi.org/10.1038/s41467-021-27045-6

Xiao H, Rasul K, Vollgraf R (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747. https://doi.org/10.48550/arXiv.1708.07747