# METRIC LEARNING IN QUANTUM FEATURE SPACES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Quantum metric learning is motivated by Hilbert space dimension, but learning depends on metric geometry: margin capacity, optimization landscapes, and evaluation tractability. By letting a classical network predict all gate parameters of a fixed ansatz, we isolate the reachable state manifold and its fidelity metric. We prove that the unentangled manifold $\mathbb{M}_{\text{unentangled}} = (\mathbb{CP}^1)^N$ saturates the $2^N$ maximal orthogonal capacity needed for classification, while entanglement adds states only at intermediate separations, which are costly to resolve under finite sampling. Global fidelity causes exponential gradient collapse across both unentangled ($\Theta(c^{-N})$) and entangled ($\Theta(4^{-N})$) manifolds; as fidelities concentrate, the fidelity signal-to-noise ratio at a fixed shot budget also decays exponentially in $N$; and under contrastive losses, unconstrained encoders reach trivial minima by *geometric shattering*. We introduce Quantum Projective Metric Learning (QPMeL), an exact $\mathcal{O}(N)$ closed-form kernel on $\mathbb{M}_{\text{unentangled}}$ that circumvents circuit simulation. QPMeL resolves gradient collapse via a curriculum annealing from a 1-local observable regime (restoring $\Theta(1/N^2)$ polynomial variance) to the global product kernel. Empirically, QPMeL matches or exceeds published quantum models across exceeds eight and matches two of 11 and confirms that depth induces kernel concentration. Execution on the 156-qubit IBM Marrakesh processor validates that physical noise heavily penalizes near-orthogonal states. Ultimately, the failures we observe in prior designs trace to kernel concentration, global fidelity objectives, and encoder regularization rather than to Hilbert-space dimension.

## 1 INTRODUCTION

Metric learning trains an encoder so that distance in an embedding space reflects semantic similarity (Chopra et al., 2005; Weinberger et al., 2005; Schroff et al., 2015), underpinning few-shot classification (Snell et al., 2017) and multimodal retrieval (Radford et al., 2021). The embedding space matters as much as the encoder: projecting onto a hypersphere bounds distances and stabilizes margin losses (Wang et al., 2017; Zhang et al., 2020). Quantum metric learning (QMeL) replaces this space with the state space of $N$ qubits (Lloyd et al., 2020; Thumwanit et al., 2021; Kim & Bekiranov, 2022; Liu et al., 2022; Hou et al., 2023), motivated by its $2^N$ dimensions (Havlíček et al., 2019; Schuld & Killoran, 2019) and quantum fidelity $F(x, x') = \left|\langle\psi(x)|\psi(x')\rangle\right|^2$, a natural positive-semidefinite kernel (Schuld, 2021). Metric learning, however, relies on metric geometry rather than raw dimension: geometric capacity at required margins, optimization landscapes, and evaluation tractability. Barren plateaus (McClean et al., 2018; Cerezo et al., 2021b), kernel concentration (Thanasilp et al., 2024), and shot noise (Miroszewski et al., 2024) govern all three, yet quantum spaces are typically judged by dimension alone. This paper evaluates them as metric-learning targets.

Evaluating quantum state spaces is obstructed in prior models because their circuits act simultaneously as kernel and function approximator (Lloyd et al., 2020; Liu et al., 2022; Hou et al., 2023), preventing failures from being attributed to the geometry or the approximator. Moreover, terminal trainable unitaries cancel identically from fidelity ($F(V|\psi\rangle, V|\phi\rangle) = F(|\psi\rangle, |\phi\rangle)$), and single-axis encodings confine qubits to 1D great circles. We decouple these roles completely: a classical network $f_w$ predicts every gate parameter of a fixed quantum ansatz $U$ with zero trainable quantum gates, $|\psi_w(x)\rangle = U(f_w(x))\,|0\rangle^{\otimes N}$ (Hur et al., 2023). All representation learning is classical; the circuit contributes solely its reachable state manifold and its fidelity metric, isolating the target quantum geometry for rigorous analysis (Section 2.1.3).



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

Our primary contributions are:

- **Formalization of Geometric Capacity Bounds and Trainability Trade-offs (Lemmas 1–2):** We prove that the unentangled manifold $\mathbb{M}_{\text{unentangled}} = (\mathbb{CP}^1)^N$ saturates the $2^N$ maximal orthogonal capacity needed for classification, while entanglement adds states only at finite separation (Lemma 1), which are costly to resolve under finite sampling (Section 2.2.2). Furthermore, we show that global fidelity objectives cause exponential gradient collapse across both unentangled ($\Theta(c^{-N})$) and entangled ($\Theta(4^{-N})$) manifolds (Lemma 2), formalizing the fundamental trade-off between state capacity ($m$) and measurement dimension ($k$).
- **Analysis of Metric Loss Functions under Finite-Shot Fidelity Estimation (FSNR & Proposition 1):** We analyze fidelity estimation under finite sampling, showing that, because fidelities concentrate, the fidelity signal-to-noise ratio (FSNR) at a fixed shot budget $S$ decays exponentially with $N$: $\Theta(S \cdot 4^{-N})$ for Haar-random states under the SWAP test and $\Theta(S(2/3)^N)$ for uniformly spread product states under the inversion test, so holding it fixed costs shots exponential in $N$. Under instance-level contrastive losses, we formalize *geometric shattering* (Proposition 1): unconstrained encoders exploit the $2^N$ orthogonal capacity of $\mathbb{M}_{\text{unentangled}}$ to achieve trivial loss minima without learning semantic representations.
- **QPMeL Regime as a Geometric Base Case with Local Training Curriculum (Section 3):** We introduce Quantum Projective Metric Learning (QPMeL), an exact $\mathcal{O}(N)$ closed-form kernel spanning $\mathbb{M}_{\text{unentangled}}$ that acts as a minimal geometric base case, training without circuit simulation or shot noise. To resolve the global fidelity barren plateau of Lemma 2, QPMeL trains via an annealing curriculum transitioning from a 1-local observable regime (which provably restores $\Theta(1/N^2)$ polynomial gradient variance) to the global product kernel.

Our empirical results across few-shot classification, multimodal EEG retrieval, and hardware execution are consistent with this analysis. QPMeL exceeds eight and matches two of eleven published few-shot baselines, and on shared EEG backbones circuit depth primarily induces kernel concentration. Because QPMeL's kernel is classically computable and the published comparisons do not match inputs or encoders, we treat these results as consistency checks rather than as evidence that restricting to $\mathbb{M}_{\text{unentangled}}$ confers an advantage. Furthermore, physical execution on the 156-qubit IBM Marrakesh processor validates our FSNR analysis: physical gate errors and shot noise disproportionately corrupt near-orthogonal negative pairs during global fidelity estimation, proving that exact classical computability avoids a fatal hardware tax. Finally, contrastive memorization occurs across shallow and deep circuits alike, confirming that representation stability requires classical inductive bias rather than quantum dimensionality. Together, these findings indicate that the failures of prior quantum metric learners trace to kernel concentration, global fidelity objectives, and classical encoder regularization rather than to Hilbert-space capacity.

## 2 Quantum Hilbert Space for Metric Learning

**Notation.** A quantum embedding maps an input $x$ to an $N$-qubit pure state $|\psi(x)\rangle \in \mathbb{CP}^{2^N-1}$ and compares two inputs by their fidelity $F$, or equivalently their Fubini–Study distance $d_{\text{FS}}$,

$$F(x,x') = \left|\langle\psi(x)|\psi(x')\rangle\right|^2 \in [0,1], \qquad d_{\text{FS}}(x,x') = \arccos\sqrt{F(x,x')} \in [0,\tfrac{\pi}{2}]. \tag{1}$$

The states available to the embedding are set by the circuit's dynamical Lie algebra (DLA) $\mathfrak{g}$, the real span of its generators and their nested commutators: the reachable states are the orbit of $|0\rangle^{\otimes N}$ under $e^{\mathfrak{g}}$. Local rotations generate $\bigoplus_q \mathfrak{su}(2)$ of dimension $3N$, with the unentangled manifold $\mathbb{M}_{\text{unentangled}} = (\mathbb{CP}^1)^N$ of product states as their orbit, and $F$ factorizes over qubits costing $\mathcal{O}(N)$ to compute (Lemma 4). Generic entangling circuits give $\mathfrak{su}(2^N)$, of dimension $4^N - 1$, whose orbit is all of $\mathbb{CP}^{2^N-1}$. Appendix A reviews metric-learning objectives, barren plateaus and kernel concentration.

### 2.1 Previous quantum metric learning

How a circuit arranges its data-dependent, trainable, and fixed gates dictates its reachable states, its fidelity metric, and the degree of classical encoder control. Prior approaches fall into three families:

#### 2.1.1 EXPLICIT VARIATIONAL QUANTUM CIRCUITS (E-VQCS)

An E-VQC applies a block $U(\vec{\theta})$ of trainable and fixed gates after a data-dependent encoding $\phi(x)$, $|\Psi_\theta(x)\rangle = U(\vec{\theta})\,\phi(x)\,|0\rangle^{\otimes N}$, as in Chen et al. (2024), Jaderberg et al. (2022) and Wang et al. (2024). This has two consequences:

**The trainable block cancels from the fidelity.** Any unitary $V$ applied after the encoding, including $V = U(\vec{\theta})$, preserves inner products (Eq. 2), so training $\vec{\theta}$ on a fidelity loss has no effect.

$$F\big(V\,|\psi(x)\rangle\,,\,V\,|\psi(x')\rangle\big) = \big|\langle\psi(x)|\,V^\dagger V\,|\psi(x')\rangle\big|^2 = F\big(|\psi(x)\rangle\,,|\psi(x')\rangle\big). \tag{2}$$

Chen et al. (2024) and Jaderberg et al. (2022) instead compare measured expectation values, so their embedding space is the vector of measurement outcomes rather than the state space.

**The encoder is never given the full manifold.** Most approaches use single-axis angle encoding, such as $R_Y(z_q)$, which limits the encoder to one great circle per qubit, wasting DNN expressivity.

#### 2.1.2 DATA RE-UPLOADING (DR)

DR repeats an encoding followed by a trainable unitary $L$ times, $|\Psi_\theta(x)\rangle = \prod_{l=1}^{L}\big[U(\theta_l)\,\phi(x)\big]\,|0\rangle^{\otimes N}$, as in Lloyd et al. (2020), Liu et al. (2022), Hou et al. (2023) and Hammad et al. (2023). Because Eq. 2 no longer applies, fidelity-based losses can be used. However:

**Access is restricted to the circuit's spectrum.** The model is a truncated Fourier series in $x$ whose frequencies are fixed by the number of re-uploads (Pérez-Salinas et al., 2020; Schuld et al., 2021), so input placement is bounded by the circuit's spectrum, which limits the access of the classical network $f_w$. Widening access with more re-uploads or richer generators costs trainability (Section A.4).

**Geometry and approximation cannot be separated.** The circuit both encodes the data and approximates the target, so a failure is hard to attribute to the geometry, to insufficient approximation capacity, or to both.

#### 2.1.3 CLASSICALLY PARAMETERIZED QUANTUM KERNELS (CPQK)

These models treat the circuit as a kernel function with no trainable parameters. For instance, Hur et al. (2023) and Liu et al. (2025) train a classical DNN $f_w$ to output the parameters of an entangling $ZZ$ feature map and optimize the fidelity directly. We consider the general form, in which the DNN emits one real parameter for every parameterized gate of a fixed ansatz:

$$z = f_w(x) = (z_1, \ldots, z_N), \quad z_q \in \mathbb{R}^L, \qquad |\psi_w(x)\rangle = U(z)\;|0\rangle^{\otimes N}\,, \tag{3}$$

for $N$ qubits and $L$ parameterized gates per qubit, with training on the kernel $k_w(x, x') = |\,\langle\psi_w(x)|\psi_w(x')\rangle\,|^2$. QPMeL (Section 3) is the base case whose access spans $\mathbb{M}_{\text{unentangled}}$. In this setup:

**The setup isolates the geometry.** The circuit defines the accessible quantum states and $f_w$ learns how to map data to them. Therefore, up to approximation error, this setup can represent the mappings of previous work whenever its circuit reaches the same output states (Appendix C.1).

**What remains is the geometry.** Because the circuit's parameters are $f_w$'s outputs, the limits of the reachable states act on the encoder directly. A kernel concentrated toward its Haar value, $\mathbb{E}[k_w] = 2^{-N}$ with $\mathrm{Var}[k_w] = \Theta(4^{-N})$ (Thanasilp et al., 2024), is a flat loss for $f_w$, and under instance-level objectives the fidelity geometry rewards *geometric shattering* (Proposition 1). Both apply to any model trained on the fidelity of the states it reaches.

### 2.2 QUANTUM HILBERT SPACE ($\mathcal{H}_Q$) AS A TARGET METRIC SPACE

In Eq. 3, the target metric space is the reachable set $\mathcal{R}(U) = \{|\psi_w(x)\rangle\}$ equipped with fidelity $F$ and distance $d_{\mathrm{FS}}$. An effective embedding space requires: (1) bounded diameter, (2) sufficient geometric capacity, and (3) a trainable, computationally tractable metric. Criterion (1) is naturally satisfied since $F \in [0, 1]$ and $d_{\mathrm{FS}} \leq \pi/2$ for all $N$, providing a fixed margin scale where exponentially many states can sit at maximal diameter (Proposition 1; see Appendices C.5–C.6 for trivial minimizers and smoothness). We now examine criteria (2) and (3).

### 2.2.1 GEOMETRIC CAPACITY

Capacity refers to how many embeddings a space can keep apart at a given resolution, measured by the maximum cardinality of an $\varepsilon$-separated set.

**Definition 1** ($\varepsilon$-Separated Sets). *A set $\mathcal{S} \subseteq \mathcal{M}$ is $\varepsilon$-separated if $d_{\mathrm{FS}}(\psi, \phi) \geq \varepsilon$ for all distinct $\psi, \phi \in \mathcal{S}$. The capacity $M(\mathcal{M}, \varepsilon)$ is the largest cardinality of such a set.*

A classification loss seeks $K$ classes at maximal separation $\varepsilon = \pi/2$, whereas retrieval tasks require graded similarity preserving a minimum gap $\varepsilon$:

**Maximal separation.** At $\varepsilon = \pi/2$, $M(\mathcal{M}, \pi/2)$ counts mutually orthogonal states. It is bounded by the dimension of the Hilbert space and attained by the computational basis, **a set that lies entirely inside $\mathbb{M}_{\text{unentangled}}$**:

$$M\big(\mathbb{M}_{\text{unentangled}}, \pi/2\big) = M\big(\mathbb{CP}^{2^N-1}, \pi/2\big) = 2^N, \tag{4}$$

so $N \geq \log_2 K$ qubits suffice: four for the ten classes of MNIST, eleven for the 1623 characters of Omniglot. Given a function approximator able to realize the assignment, neither entanglement nor further qubits add capacity at this separation. Under an *instance-level* objective such as InfoNCE, *where each sample is its own class*, the same capacity works against the embedding (Appendix E.2).

**Proposition 1** (Geometric shattering). *(i) On $\mathbb{M}_{\textit{unentangled}}$, $F = \prod_q F_q$ is zero if any one qubit pair is orthogonal, and one only if all are aligned. (ii) With fidelity similarity at temperature $\tau$, InfoNCE (Eq. B.3) on $B$ pairs is at least $\log\big(1 + (B-1)e^{-1/\tau}\big)$, with equality exactly at $k_w(x_i, y_j) = \delta_{ij}$, which $\mathbb{M}_{\textit{unentangled}}$ attains, by any assignment of samples to orthonormal states, if and only if $B \leq 2^N$.*

At such a minimum the gradients become data-independent and are not meaningful outside the training set; we call this *geometric shattering*. InfoNCE rewards raising a positive and lowering a negative equally, but lowering is easier: one orthogonal qubit sets the fidelity to zero, whereas raising it requires aligning every qubit through multiplicative gradients (Eq. 9). Under standard cosine similarity, a classical unit sphere cannot reach this floor for $B > 2$ since no three vectors can be pairwise antipodal. On $\mathbb{M}_{\text{unentangled}}$, however, the metric reaches its minimum at orthogonality, which $2N$ parameters attain for any $B \leq 2^N$ (as would a classical product kernel on $(S^2)^N$). Preventing shattering is therefore the job of $f_w$, through a Lipschitz bound (Kim & Bekiranov, 2022) or dropout (Section 4.3).

**Finite resolution.** At intermediate $\varepsilon$, capacity is a packing problem. Comparing an $\varepsilon$-separated set with the fraction of the space a Fubini–Study ball (Definition 3) occupies gives the bounds:

**Lemma 1** (Capacity at finite resolution). *Let $m = \dim_{\mathbb{R}}(\mathcal{M})$ be the **real dimension** of the manifold, with logarithms taken base two so that $\log M$ counts bits. For fixed $\varepsilon \in (0, \pi/2)$,*

$$m \log_2 \frac{1}{\sin \varepsilon} \;\leq\; \log_2 M(\mathcal{M}, \varepsilon) \;\leq\; m \log_2 \frac{1}{\sin(\varepsilon/2)} + \Delta, \tag{5}$$

*with $\Delta = 0$ for $\mathcal{M} = \mathbb{CP}^{2^N-1}$, where $m = 2(2^N - 1)$, and $\Delta = \mathcal{O}(N \log N)$ for $\mathcal{M} = \mathbb{M}_{\textit{unentangled}}$, where $m = 2N$. (Proof in Appendix C.2.)*

The log-capacity thus grows with the real dimension of the manifold. At $\varepsilon = \pi/4$, where two states still have fidelity $\frac{1}{2}$, the full space holds at least $2^{2^N-1}$ states, while the product manifold holds between $6^N$ and $(4 + 2\sqrt{2})^N N^N$ ($2^{\Theta(N \log N)}$ vs. $2^{\Theta(2^N)}$). Entanglement therefore adds states at intermediate separations, not at maximal separation, and the closer together those states lie, the more shots it takes to tell them apart: resolving separation $\varepsilon$ costs $S \in \Omega(\varepsilon^{-2})$ shots (scaling to $\Omega(\varepsilon^{-4})$ under SWAP tests for near-orthogonal pairs; Section 2.2.2). A model can use the added states only as far as its shot budget resolves $\varepsilon$, and under concentration most pairs end up nearly orthogonal.

**Reachable states.** Every state an embedding uses must be reachable. The reachable set is the orbit of $|0\rangle^{\otimes N}$ under $e^{\mathfrak{g}}$, of dimension at most $\dim \mathfrak{g}$, so occupying $m$ real dimensions requires $\dim \mathfrak{g} \geq m$. Local rotations reach all of $\mathbb{M}_{\text{unentangled}}$ with $\dim \mathfrak{g} = 3N$. Covering $\mathbb{CP}^{2^N-1}$ requires an algebra that acts transitively on it, and generic entangling ansätze close to $\mathfrak{su}(2^N)$, for which

$$\dim \mathfrak{g} \;=\; 4^N - 1 \;=\; \left(\frac{m}{2}\right)^2 + m, \tag{6}$$

quadratic in the dimension reached: all but $m$ of these directions leave the state unchanged (proof in Appendix C.3).

**Gradient collapse under global fidelity.** Metric learning evaluates pairwise similarity via fidelity $F(x, x') = \left|\langle\psi(x)|\psi(x')\rangle\right|^2$, which measures a global rank-1 observable $O_G = |\psi(x')\rangle\langle\psi(x')|$ across all $N$ qubits ($2^N$ dimensions). As proved by Cerezo et al. (2021b), **global cost functions suffer from an exponential barren plateau at random initialization, with or without entanglement**:

**Lemma 2** (Gradient collapse under global fidelity)**.** *Let an $N$-qubit embedding circuit have parameterization $\theta$, and let $C_G(\theta) = 1 - |\langle\psi(\theta)|\phi\rangle|^2$ be a global fidelity cost against a product target $|\phi\rangle$. On $\mathbb{M}_{unentangled}$, let the per-qubit parameters be independent and uniformly distributed; on $\mathbb{CP}^{2^N-1}$, let the differentiated gate sit between two independent unitary 2-designs over $\mathrm{SU}(2^N)$. Then the gradient variance vanishes exponentially with $N$ on both manifolds:*

$$\mathrm{Var}_\theta[\partial_\mu C_G] = \Theta(c^{-N}) \text{ on } \mathbb{M}_{unentangled}, \qquad \mathrm{Var}_\theta[\partial_\mu C_G] = \Theta(4^{-N}) \text{ on } \mathbb{CP}^{2^N-1}, \tag{7}$$

*where $c = 1/\mathbb{E}[F_q^2] > 1$ depends on the angle distribution and the target: $c = 3$ for Haar-distributed single-qubit states and $c = 8/3$ for single-axis rotations against $|0\rangle^{\otimes N}$. Moreover, the parameter volume with $C_G \le \delta$ vanishes as $\mathcal{O}((1/2)^N)$ (Cerezo et al., 2021b). (Proof in Appendix E.1.)*

Lemma 2 averages over uniformly random parameters, so it describes initialization. When the angles are outputs of $f_w$, the relevant distribution is the one $f_w$ induces, and once positives are aligned ($F_q \approx 1$ on every qubit) the bound no longer applies. This is why QPMeL can start on a local cost and end on the global kernel (Section 3.3).

**The tradeoff: State capacity vs. measurement dimension.** Trainability is governed not merely by the manifold, but by the number of dimensions in which fidelity is measured. Measuring fidelity locally on a $k$-qubit subsystem ($k \ll N$) yields polynomial gradient variance $\Omega(1/\mathrm{poly}(N))$ for shallow ($\mathcal{O}(\log N)$-depth) circuits and broadens the optimization gorge (Cerezo et al., 2021b). However, distinguishing exponentially many states or exploiting quantum correlations requires making the measurement global ($k \to N$), which incurs the cost of exponential gradient collapse (Eq. 7). A local cost can be defined for any embedding, entangled or not (Cerezo et al., 2021b); on $\mathbb{M}_{\text{unentangled}}$ the 1-local cost is additionally computable exactly and classically (Section 3.3).

### 2.2.2 The fidelity inner product

Fidelity is a promising similarity: it is a bounded positive-semidefinite (PSD) kernel and can be estimated on hardware for classically intractable states. Three costs, however, grow with the qubit count $N$.

**Fidelity is estimated, not computed.** Unless the state is classically tractable, resolving a fidelity gap $\delta$ costs $S \in \Omega(\delta^{-2})$ shots; under the SWAP test on near-orthogonal pairs ($F \approx 0$), resolving separation $\varepsilon$ costs $S \in \Omega(\varepsilon^{-4})$ shots, resolving only $\varepsilon \gtrsim S^{-1/4}$ in that regime. The fidelity signal-to-noise ratio (FSNR) quantifies this:

$$\mathrm{FSNR} \;\equiv\; \frac{\mathrm{Var}_{x,x'}\left[F(x,x')\right]}{\mathbb{E}_{x,x'}\left[\mathrm{Var}(\hat{F} \mid F)\right]}, \tag{8}$$

FSNR is $S \cdot 2^{-N}$ for Haar-random states using the inversion test, and $\Theta(S \cdot 4^{-N})$ using the SWAP test. For uniformly spread product states under the inversion test it is $\Theta\big(S(2/3)^N\big)$ ($\Theta(S \cdot 3^{-N})$ under SWAP; Appendix E.3), so holding it fixed costs shots exponential in $N$ across all cases. Hardware noise also biases the estimate, and this error does not shrink as $S \to \infty$ (Wang et al., 2021).

**The gradient is multiplicative.** On the product manifold the kernel factorizes, and so does its derivative,

$$F = \prod_q F_q \qquad \Longrightarrow \qquad \frac{\partial F}{\partial F_q} = \prod_{q' \neq q} F_{q'}. \tag{9}$$

One qubit at $F_{q'} = 0$ zeroes the gradient of every other, and at $F_q \approx 1/2$ the per-qubit gradient is $2^{-(N-1)}$, or $5 \times 10^{-4}$ at $N = 12$. This multiplicative decay is the product-state manifestation of the global fidelity barren plateau (Cerezo et al., 2021b), arising with zero entanglement and constant depth. Section 3 resolves this by transitioning from local to global fidelity.

**Every pair costs a circuit.** A generic state has no efficient classical description, so each pair in a batch needs its own execution: $5 \times 10^{10}$ circuit executions per step for a unimodal batch of $10^4$ at $10^3$ shots ($10^{11}$ for cross-modal), and a query costs $\mathcal{O}(|\mathrm{DB}|)$ runs. On $\mathbb{M}_{\text{unentangled}}$, Lemma 4 returns $F$ in $\mathcal{O}(N)$. For small misalignments $F$ is a Gaussian kernel whose bandwidth $f_w$ sets, so shattering is a kernel that is too narrow (Appendix C.4).

### 2.3 Design implications

**Kernel metric learning is the most promising.** The metric is fixed by the geometry of the reachable states, and the expressivity of $f_w$ is cheap by comparison (Eq. 3).

**Train on local fidelity before global fidelity.** Global fidelity causes barren plateaus and shot bottlenecks across both unentangled and entangled spaces at random initialization (Lemma 2). A local surrogate keeps gradients polynomial while the encoder leaves this regime; on $\mathbb{M}_{\text{unentangled}}$ the surrogate is exact and classically computable, and annealing to the global kernel recovers its $2^N$ orthogonal states (Section 3.3).

**Design the encoder for the geometry.** Concentration flattens the loss, and shattering is $f_w$ spending the orthogonal states the geometry allows.

## 3 Quantum Projective Metric Learning (QPMeL)

QPMeL instantiates Eq. 3 with the smallest ansatz that reaches all of $\mathbb{M}_{\text{unentangled}}$, $U = \bigotimes_q R_Z(\gamma_q) R_Y(2\theta_q)$, so $L = 2$: two angles per qubit. These rotations generate $\bigoplus_q \mathfrak{su}(2)$ of dimension $3N$ (versus $4^N - 1$ for generic entangling circuits). Its kernel is closed form (Lemma 4) and training is classical, avoiding variational parameter-shift and shot noise. Since E-VQC and re-uploading reach entangled states, $\mathbb{M}_{\text{unentangled}}$ is a submanifold of what they prepare; QPMeL serves as the unentangled baseline.

### 3.1 Split encoder and angular rescaling

The encoder is split into a domain model and an angle head (Figure D.1). The domain model $h : \mathcal{X} \to \mathbb{R}^d$ is chosen for the modality (a CNN, residual MLP or transformer) and does all of the representation learning. The angle head maps $h(x)$ to $N$ coordinate pairs through bounded activations,

$$\theta_q(x) = \pi \,\mathrm{sigmoid}\big(w_{\theta,q}^\top h(x) + b_{\theta,q}\big) \in (0, \pi), \qquad \gamma_q(x) = \pi \tanh\big(w_{\gamma,q}^\top h(x) + b_{\gamma,q}\big) \in (-\pi, \pi). \tag{10}$$

Because $R_Y$ and $R_Z$ are $2\pi$-periodic, unbounded outputs would map distant inputs to the same state; the bounded activations keep each angle within one period. The resulting state is

$$|\psi(x)\rangle = \bigotimes_{q=1}^{N} \Big( \cos\theta_q(x)\,|0\rangle + e^{i\gamma_q(x)} \sin\theta_q(x)\,|1\rangle \Big) \tag{11}$$

### 3.2 The projective metric function

Training never runs the circuit (Figure D.2). Instead it evaluates the fidelity of Eq. 11 classically with the *projective metric function* (PMeF), a two-step map from angles to fidelity. The first step maps each qubit's angles to a real unit vector,

$$C_q = (x_q, y_q, z_q) = (\sin\theta_q \cos\gamma_q,\ \sin\theta_q \sin\gamma_q,\ \cos\theta_q), \tag{12}$$

in which Eq. 11 reads $\bigotimes_q \big(z_q\,|0\rangle + (x_q + i y_q)\,|1\rangle\big)$. The second step computes the fidelity from these vectors alone, since the single-qubit overlap $\langle\psi_q|\phi_q\rangle = \lambda_q^r + i\lambda_q^c$ splits into a real and an imaginary part,

$$\lambda_q^r = x_q x_q' + y_q y_q' + z_q z_q', \qquad \lambda_q^c = x_q y_q' - y_q x_q'. \tag{13}$$

**Definition 2** (Complex Kernel Function)**.** *For single-qubit states $|\psi_q\rangle, |\phi_q\rangle$ with coordinate vectors $C_q$ and $C_q'$,* $\mathrm{CKF}(C_q, C_q') = (\lambda_q^r)^2 + (\lambda_q^c)^2 = |\langle\psi_q|\phi_q\rangle|^2$.

By Lemma 4 the $N$-qubit fidelity is the product of these factors, and composing the two steps gives the model's kernel,

$$k_w(x, x') = \mathrm{PMeF}(x, x') = \prod_{q=1}^{N} F_q, \qquad F_q = \mathrm{CKF}\big(C_q(x), C_q(x')\big), \tag{14}$$

evaluated in $\mathcal{O}(N)$ real arithmetic per pair. Each factor equals $\frac{1}{2}(1 + \vec{b}_q \cdot \vec{b}'_q)$ for the Bloch vector $\vec{b}_q = (\sin 2\theta_q \cos\gamma_q, \sin 2\theta_q \sin\gamma_q, \cos 2\theta_q)$, an inner product of real features, so the kernel is positive semi-definite by the Schur product theorem (Appendix E.6). Training thus avoids circuit simulation, parameter-shift gradients and the shot budget of Section 2.2.2, and its coordinates prepare the state of Eq. 11 directly.

### 3.3 Local observable training and the additive-to-product curriculum

Section 2.2 established that measuring fidelity globally across all $N$ qubits collapses gradient variance exponentially ($\Theta(c^{-N})$, Lemma 2), while measuring fidelity in a local observable regime ($k \ll N$) yields polynomial trainability for shallow circuits (Cerezo et al., 2021b). Because states on $\mathbb{M}_{\text{unentangled}}$ factorize over qubits, the 1-local cost surrogate ($k = 1$) is computed exactly and classically: the *additive kernel* $\bar{k}_w$, which averages the single-qubit fidelity factors rather than multiplying them,

$$\bar{k}_w(x, x') = \frac{1}{N}\sum_{q=1}^{N} F_q, \qquad \frac{\partial \bar{k}_w}{\partial \theta_q} = \frac{1}{N}\frac{\partial F_q}{\partial \theta_q}. \tag{15}$$

This evaluates the 1-local observable $O_L = \frac{1}{N}\sum_q |\psi'_q\rangle\langle\psi'_q| \otimes \mathbb{I}_{\bar{q}}$, yielding provably polynomial gradient variance $\mathrm{Var}[\partial \bar{k}_w/\partial\theta_q] = \Theta(1/N^2)$ (Appendix E.1) and escaping the narrow gorge of global fidelity.

However, pure local training has limited margin: $\bar{k}_w$ requires all $N$ qubits to be orthogonal to reach zero similarity, whereas the global product kernel $k_w$ is minimized as soon as a single qubit pair is orthogonal. To navigate the tradeoff between local trainability and global separation advantage, QPMeL anneals linearly from the local observable to the global product kernel over $T$ epochs,

$$S_t(x, x') = \alpha_t\, k_w(x, x') + (1 - \alpha_t)\, \bar{k}_w(x, x'), \qquad \alpha_t = t/T. \tag{16}$$

The two kernels agree on positives ($F = 1$ only when all qubits align). The curriculum begins in the local observable regime where gradients are distributed across every qubit independently, and gradually transitions to the global product fidelity, ending on the exact PMeF of Eq. 14. Lemma 2 bounds the global kernel's gradients only at random initialization; by the time the curriculum reaches it, positives are aligned and the bound no longer applies.

QPMeL gives the encoder control of all $2N$ real dimensions of $\mathbb{M}_{\text{unentangled}}$ and nothing beyond it. A single rotation axis gives the encoder half of these dimensions, and a circuit that reaches beyond $\mathbb{M}_{\text{unentangled}}$ can concentrate the kernel or lose input structure (Section 4.3); the experiments compare these designs but do not isolate reachable dimension from the other ways they differ.

## 4 Results

**Setup.** QPMeL reaches exactly $\mathbb{M}_{\text{unentangled}}$ and nothing beyond it (Section 3), so each comparison below asks whether a model's additional state access helps the classical encoder attached to it. MNIST (Deng, 2012) and Fashion-MNIST (Xiao et al., 2017) use a five-layer CNN with 5 and 6 qubits, Omniglot (Lake et al., 2015) a three-block CNN with 11, and THINGS-EEG2 (Gifford et al., 2022) the EEG encoders of Chen et al. (2024) against frozen CLIP image features (Radford et al., 2021) with 10. Few-shot tasks train a class-level objective and classify each query by its nearest class prototype; retrieval tasks train the symmetric InfoNCE objective of CLIP. The similarity is always the PMeF of Eq. 14, computed classically except in Section 4.4.

### 4.1 Few-shot classification

$\mathbb{M}_{\text{unentangled}}$ holds $2^N$ mutually orthogonal states (Eq. 4), so a $K$-class task on $N \geq \log_2 K$ qubits should lose nothing to circuits that reach further. Table 1 tests this against quantum models from five

papers, each on its own protocol. *Circuit* is the family of Section 2.1, qubit counts are in parentheses, and bold marks a difference beyond QPMeL's 95% confidence interval.

QPMeL exceeds eight of the eleven baselines, matches two and trails one, with fewer or equal qubits in ten of the eleven comparisons. The largest margins are at one shot, where the strongest models of Wang et al. (2024) reach $50.5\%$ on 3-way MNIST and $62.2\%$ on 3-way Fashion-MNIST against $98.3\%$ and $94.0\%$. Every baseline reaches beyond $\mathbb{M}_{\text{unentangled}}$ with entangling or trainable layers, and none of that added state access appears as accuracy: QPMeL classifies all ten MNIST and Fashion-MNIST classes with 5 and 6 qubits, close to the $\lceil \log_2 10 \rceil = 4$ that Eq. 4 requires. The one model it trails, a CNN feeding a 64-qubit re-uploading circuit (Liu et al., 2022), uses nearly six times as many qubits (Appendix F.1).

### 4.2 Multimodal retrieval

**THINGS-EEG2.** Table 2 keeps the EEG encoders of Chen et al. (2024), convolutional (QSTConv) and transformer (QNervFormer), bit-identical and changes only the head: their E-VQC with a $\langle Z \rangle$ cosine readout, the same E-VQC and its re-uploading variant (DR) at $L$ layers, both scored by fidelity, and QPMeL. *Train* is 200-way retrieval among 200 training images and *held-out* the 200-way zero-shot task on unseen concepts; chance is $0.5\%$. Over 10 subjects and 5 seeds on QSTConv, QPMeL reaches $23.96 \pm 5.2\%$ held-out against $0.84 \pm 0.8\%$ for the E-VQC. Our E-VQC matches the equations of Chen et al. (2024) but not their accuracy, so we compare heads only on shared backbones (Appendix F.2).

### 4.3 Trainability analysis

We test the two CPQK failures of Section 2.1.3, concentration and shattering (details in Appendix F.3).

**Depth concentrates the kernel.** DR's last CNOT ring cancels (Eq. 2), so $L$ layers leave $L-1$ rings in the kernel, and each ring removes input structure: at initialization the kernel's rank correlation $\rho$ with input similarity falls from $0.64$ at one layer to $0.46$, $0.06$ and $0$ at two, four and eight, where its mean and spread match Haar-random states. Held-out accuracy falls with each ring while the first-epoch encoder gradient grows nearly 500-fold from one layer to sixteen, so the kernel fails, not the gradient.

**QPMeL controls all of $\mathbb{M}_{\text{unentangled}}$.** Without entangling rings, 1-layer DR reduces to a product kernel on $\mathbb{M}_{\text{unentangled}}$ with per-qubit factors $1 - \sin^2 w_q \sin^2((x_q - x'_q)/2)$. Its trainable weights $w_q$ adjust contrast, so it outperforms the E-VQC, whose layers cancel, but trails QPMeL ($12.3\%$ vs. $19.9\%$ on QSTConv; $6.6\%$ vs. $11.8\%$ with branch dropout). Its independent polar angles start the

Table 1: *Few-shot classification against published quantum models.*

| Model | Circuit | Protocol | Baseline | **QPMeL** |
|---|---|---|---|---|
| | | ***MNIST*** | | |
| Thumwanit et al. (2021) | E-VQC | Binary $(3, 6)$ | 91.7 (9) | $\mathbf{99.80 \pm 0.20}$ (5) |
| NQE (Hur et al., 2023) | CPQK | Binary $(0, 1)$ | 99.50 (12) | $99.60 \pm 1.71$ (5) |
| Huang et al. (2025) | E-VQC | 2w-10s $(3, 5)$ | 89.69 (9) | $\mathbf{99.00 \pm 1.01}$ (5) |
| Wang et al. (2024) | E-VQC | 3w-1s $(0, 1, 2)$ | 50.5 (6) | $\mathbf{98.27 \pm 1.86}$ (5) |
| Wang et al. (2024) | E-VQC | 3w-10s $(0, 1, 2)$ | 82.5 (6) | $\mathbf{99.73 \pm 0.37}$ (5) |
| No Baseline | | 10w-1s (0–9) | – | $96.7 \pm 0.6$ (5) |
| | | ***Fashion-MNIST*** | | |
| NQE (Hur et al., 2023) | CPQK | Binary $(0, 1)$ | $95.5 \pm 0.41$ (12) | $\mathbf{99.60 \pm 0.95}$ (6) |
| Huang et al. (2025) | E-VQC | 2w-10s $(0, 1)$ | 96.92 (9) | $97.80 \pm 1.53$ (6) |
| Wang et al. (2024) | E-VQC | 3w-1s $(0, 1, 2)$ | 62.2 (6) | $\mathbf{94.00 \pm 2.99}$ (6) |
| Wang et al. (2024) | E-VQC | 3w-10s $(0, 1, 2)$ | 73.5 (6) | $\mathbf{96.67 \pm 1.44}$ (6) |
| No Baseline | | 10w-1s (0–9) | – | $85.03 \pm 0.38$ (6) |
| | | ***Omniglot*** | | |
| Huang et al. (2025) | E-VQC | 2w-10s $(0, 1)$ | 93.83 (9) | $\mathbf{98.00 \pm 0.85}$ (11) |
| Liu et al. (2022) | DR | 5w-5s | **98.0** (64) | $96.16 \pm 0.70$ (11) |
| No Baseline | | 10w-5s | – | $90.02 \pm 0.42$ (11) |

Table 2: *THINGS-EEG2 200-way top-1 accuracy (%) with only the head changed; $L$ circuit layers.*

| | | | QSTConv | | QNervFormer | | QNervFormer (dropout) | |
|---|---|---|---|---|---|---|---|---|
| Head | Similarity | $L$ | Train | Held-out | Train | Held-out | Train | Held-out |
| Chen et al. | $\langle Z \rangle$ cosine | 4 | – | 3.2 | – | 4.3 | – | – |
| E-VQC | Fidelity | 4 | $1.5 \pm 0.9$ | $1.2 \pm 0.4$ | $10.1 \pm 2.6$ | $0.2 \pm 0.3$ | $2.8 \pm 0.4$ | $2.3 \pm 1.4$ |
| DR | Fidelity | 1 | $13.1 \pm 6.5$ | $12.3 \pm 5.2$ | $65.0 \pm 17.9$ | $1.0 \pm 0.9$ | $7.9 \pm 6.2$ | $6.6 \pm 4.6$ |
| DR | Fidelity | 2 | $2.9 \pm 1.8$ | $3.7 \pm 1.2$ | $43.0 \pm 8.0$ | $0.3 \pm 0.6$ | $2.9 \pm 2.4$ | $2.3 \pm 2.0$ |
| DR | Fidelity | 4 | $0.6 \pm 0.7$ | $0.7 \pm 0.4$ | $6.5 \pm 1.7$ | $0.5 \pm 0.4$ | $0.9 \pm 1.0$ | $0.9 \pm 0.7$ |
| **QPMeL** | PMeF | – | $\mathbf{28.0 \pm 3.4}$ | $\mathbf{19.9 \pm 3.1}$ | $\mathbf{96.5 \pm 0.8}$ | $1.1 \pm 0.4$ | $\mathbf{21.2 \pm 8.6}$ | $\mathbf{11.8 \pm 5.6}$ |

kernel concentrated with a multiplicative gradient (Eq. 9), and runs linger on the $\ln 1024$ loss plateau for 30–125 epochs. QPMeL's additive start instead gives 1.39 nats of initial range and its loss falls from epoch 1; ending on the product kernel adds $\sim$5 points of held-out top-1 on QSTConv over a purely additive kernel (Table F.1).

**The encoder memorizes; geometry decides whether it shatters.** Under QPMeL, the unregularized transformer memorizes ($96.5\%$ train vs. $1.1\%$ held-out), as it does under an additive kernel that can orthogonalize at most two states (Table F.1), so memorization originates in the encoder, while the product geometry induces shattering. With a batch of $2^{10} \leq 2^N$, memorization barely changes matched fidelity ($0.50 \to 0.56$) but cuts unmatched fidelity fivefold ($0.25 \to 0.05$), with $33\%$ of negatives falling below $0.01$ (the cheap direction of Proposition 1). Branch dropout on the encoder restores generalization ($0.8\% \to 11.8\%$). Shattering thus requires both an unconstrained encoder and a batch within the $2^N$ orthogonal capacity.

### 4.4 Hardware execution and noisy simulation

On $\mathbb{M}_{\text{unentangled}}$ the fidelity is exact in closed form, so hardware execution isolates the estimation costs of Section 2.2.2. Table 3 runs the MNIST, Fashion-MNIST and Omniglot encoders on the 156-qubit IBM Marrakesh processor with SWAP tests (73–184 routed two-qubit gates), and simulates inversion tests under an IBM-Prague noise model without shot noise (Appendix F.5). Under simulated noise, MNIST and Omniglot stay within 0.5 points of the exact kernel and Fashion-MNIST loses 4.1. On hardware, MNIST keeps its accuracy within error, while Fashion-MNIST and Omniglot lose about 10 points. The SWAP estimate's variance, $(1 - F^2)/S$, exceeds the inversion test's $F(1 - F)/S$ by $(1 + F)/F$, penalizing near-orthogonal negatives most. The closed form on $\mathbb{M}_{\text{unentangled}}$ avoids these estimation costs entirely.

## 5 Conclusion

The dimension of the quantum Hilbert space is a poor measure of its usefulness for metric learning. Separating the circuit from the function approximator shows that the unentangled manifold already holds the $2^N$ orthogonal states that classification needs, and that entanglement adds states only at intermediate separations, where concentration makes fidelity costly to estimate. Global fidelity has a barren plateau at random initialization on both manifolds, trainable blocks placed after the encoding cancel from the kernel, and under contrastive losses the same orthogonal capacity lets an unconstrained encoder shatter the batch. Metric learning in quantum feature spaces is therefore limited by kernel concentration, the choice of fidelity objective and encoder regularization rather than by the size of the state space; QPMeL and our experiments are consistent with this picture.

Table 3: *Few-shot accuracy (%) on IBM hardware and under a device noise model.*

| Dataset | Task | SWAP 2q gates | Shots | Exact | IBM-Prague model | IBM Marrakesh |
|---|---|---|---|---|---|---|
| MNIST | 10w-1s | 73 | 128 | $96.7 \pm 0.6$ | $96.3 \pm 0.8$ | $97.6 \pm 3.4$ |
| Fashion-MNIST | 10w-1s | 93 | 128 | $85.03 \pm 0.4$ | $80.9 \pm 1.5$ | $74.8 \pm 4.2$ |
| Omniglot | 5w-3s | 184 | 128 | $93.9 \pm 0.7$ | $93.4 \pm 1.2$ | $84.0 \pm 6.0$ |

### AI USE STATEMENT

We used AI to assist with coding and automating experiments. AI was also used for writing assistance for example grammer passes and making language more compact. Additionally, we used AI to generate reviews and feedback to polish our work and catch mistakes. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This paper does not involve human subjects, data sets, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, or research integrity issues.

### REPRODUCIBILITY STATEMENT

This paper uses publicly available data sets and the source code will be made available to the community on github under a Apache 2.0 license.

## REFERENCES

Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. The curse of highly variable functions for local kernel machines. In *Advances in Neural Information Processing Systems*, volume 18, pp. 107–114, 2005.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, August 2021a. ISSN 2522-5820. doi: 10.1038/s42254-021-00348-9. URL http://dx.doi.org/10.1038/s42254-021-00348-9.

M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1):1791, 2021b. doi: 10.1038/s41467-021-21728-w.

Chi-Sheng Chen, Aidan Hung-Wen Tsai, and Sheng-Chieh Huang. Quantum multimodal contrastive learning framework. *arXiv preprint arXiv:2408.13919*, 2024.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. URL https://arxiv.org/abs/1504.00325.

François Chollet. Xception: Deep learning with depthwise separable convolutions, 2017. URL https://arxiv.org/abs/1610.02357.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546 vol. 1, 2005. doi: 10.1109/CVPR.2005.202.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012.

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz S Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. doi: 10.1016/j.neuroimage.2022.119754.

A. Hammad, Kyoungchul Kong, Myeonghun Park, and Soyoung Shim. Quantum metric learning for new physics searches at the LHC. *arXiv preprint arXiv:2311.16866*, 2023.

Aram W. Harrow and Richard A. Low. Random quantum circuits are approximate 2-designs. *Communications in Mathematical Physics*, 291(1):257–302, 2009.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, mar 2019. doi: 10.1038/s41586-019-0980-2. URL https://doi.org/10.1038%2Fs41586-019-0980-2.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network, 2018.

Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J. Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3(1):010313, 2022.

Yan-Yan Hou, Jian Li, Xiu-Bo Chen, and Chong-Qiang Ye. Quantum adversarial metric learning model based on triplet loss function. *EPJ Quantum Technology*, 10(1):1–16, December 2023. ISSN 2196-0763. doi: 10.1140/epjqt/s40507-023-00182-1. URL https://epjquantumtechnology.springeropen.com/articles/10.1140/epjqt/s40507-023-00182-1. Number: 1 Publisher: SpringerOpen.

Zhihao Huang, Jinjing Shi, and Xuelong Li. Quantum few-shot image classification. *IEEE Transactions on Cybernetics*, 55(1):194–206, 2025. doi: 10.1109/TCYB.2024.3476339.

Thomas Hubregtsen, David Wierichs, Elies Stecher, and Johannes Jakob Meyer. Training quantum embedding kernels on near-term quantum computers. *Physical Review A*, 106(4):042431, 2022.

Tak Hur, Israel F. Araujo, and Daniel K. Park. Neural quantum embedding: Pushing the limits of quantum supervised learning, 2023. URL https://doi.org/10.1103/PhysRevA.110.022411.

Ben Jaderberg, Lewis W. Anderson, Weidi Xie, Samuel Albanie, Martin Kiffner, and Dieter Jaksch. Quantum self-supervised learning, 2022. URL https://arxiv.org/abs/2103.14653.

Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. Quantum machine learning beyond kernel methods. *Nature Communications*, 14(1), jan 2023. doi: 10.1038/s41467-023-36159-y. URL https://doi.org/10.1038%2Fs41467-023-36159-y.

Jonathan Kim and Stefan Bekiranov. Generalization performance of quantum metric learning classifiers. *Biomolecules*, 12(11):1576, 2022.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050. URL https://www.science.org/doi/abs/10.1126/science.aab3050.

Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo. Diagnosing barren plateaus with tools from quantum optimal control. *Quantum*, 6:824, 2022.

Ryan LaRose and Brian Coyle. Robust data encodings for quantum classifiers. *Physical Review A*, 102(3), sep 2020. ISSN 2469-9934. doi: 10.1103/physreva.102.032420. URL http://dx.doi.org/10.1103/PhysRevA.102.032420.

Linh Le and Ying Xie. Deep embedding kernel, 2018.

Hongfeng Liu, Tak Hur, and Shitao Zhang. Neural quantum embedding via deterministic quantum computation with one qubit. *arXiv preprint arXiv:2501.15359*, 2025.

Minzhao Liu, Junyu Liu, Rui Liu, Henry Makhanov, Danylo Lykov, Anuj Apte, and Yuri Alexeev. Embedding learning in hybrid quantum-classical neural networks. In *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, sep 2022. doi: 10.1109/qce53715.2022.00026. URL https://doi.org/10.1109%2Fqce53715.2022.00026.

Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning, 2020. URL https://doi.org/10.48550/arXiv.2001.03622.

Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9 (1), nov 2018. doi: 10.1038/s41467-018-07090-4. URL https://doi.org/10.1038%2Fs41467-018-07090-4.

Artur Miroszewski, Marco Fellous Asiani, Jakub Mielczarek, Bertrand Le Saux, and Jakub Nalepa. In search of quantum advantage: Estimating the number of shots in quantum kernel methods, 2024. URL https://arxiv.org/abs/2407.15776.

K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Physical Review A*, 98(3), September 2018. ISSN 2469-9934. doi: 10.1103/physreva.98.032309. URL http://dx.doi.org/10.1103/PhysRevA.98.032309.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, February 2020. ISSN 2521-327X. doi: 10.22331/q-2020-02-06-226. URL http://dx.doi.org/10.22331/q-2020-02-06-226.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Michael Ragone, Bojko N. Bakalov, Frédéric Sauvage, Alexander F. Kemper, Carlos Ortiz Marrero, Martín Larocca, and M. Cerezo. A lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nature Communications*, 15(1), August 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49909-3. URL http://dx.doi.org/10.1038/s41467-024-49909-3.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. doi: 10.1109/cvpr.2015.7298682. URL http://dx.doi.org/10.1109/CVPR.2015.7298682.

Maria Schuld. Supervised quantum machine learning models are kernel methods, 2021.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4), feb 2019. doi: 10.1103/physrevlett.122.040504. URL https://doi.org/10.1103%2Fphysrevlett.122.040504.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3), March 2021. ISSN 2469-9934. doi: 10.1103/physreva.103.032430. URL http://dx.doi.org/10.1103/PhysRevA.103.032430.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017. URL https://arxiv.org/abs/1703.05175.

Supanut Thanasilp, Samson Wang, M. Cerezo, and Zoë Holmes. Exponential concentration in quantum kernel methods. *arXiv preprint arXiv:2208.11060*, 2024.

Napat Thumwanit, Chayaphol Lortaraprasert, Hiroshi Yano, and Rudy Raymond. Trainable discrete feature embeddings for variational quantum classifier, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss, 2017.

Ruhan Wang, Ye Wang, Jing Liu, and Toshiaki Koike-Akino. Quantum diffusion models for few-shot learning. *arXiv preprint arXiv:2411.04217*, 2024.

Xinbiao Wang, Yuxuan Du, Yong Luo, and Dacheng Tao. Towards understanding the power of quantum kernels in the nisq era. *Quantum*, 5:531, August 2021. ISSN 2521-327X. doi: 10.22331/q-2021-08-30-531. URL http://dx.doi.org/10.22331/q-2021-08-30-531.

Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/a7f592cef8b130a6967a90617db5681b-Paper.pdf.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision (ECCV)*, 2016.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Dingyi Zhang, Yingming Li, and Zhongfei Zhang. Deep metric learning with spherical embedding, 2020.

# A BACKGROUND

**A.1 Metric learning.** A metric-learning model has three parts: an encoder $f_w : \mathcal{X} \to \mathcal{M}$ with parameters $w$, a target space $\mathcal{M}$ with a distance $d$ or similarity $s$, and a loss that pulls similar inputs together and pushes dissimilar ones (negatives) apart by a margin (Appendix B.1). Quantum metric learning takes $\mathcal{M}$ to be the state space of $N$ qubits and $s$ to be the fidelity, which makes the model a quantum kernel method (Appendix B.3).

**A.2 The quantum state space.** Pure states of $N$ qubits are unit vectors in $\mathcal{H} = (\mathbb{C}^2)^{\otimes N}$, a $2^N$-dimensional Hilbert space; identifying vectors that differ by a global phase gives the complex projective space $\mathbb{CP}^{2^N-1}$. Its similarity and distance are the fidelity $F$ and the Fubini–Study distance $d_{\mathrm{FS}}$ of Eq. 1. Since $F = \mathrm{Tr}[\rho\rho']$ is an inner product of the density matrices $\rho = |\psi\rangle\langle\psi|$, it is a bounded positive-semidefinite (PSD) kernel, and $d_{\mathrm{FS}}$ makes $\mathbb{CP}^{2^N-1}$ a metric space of diameter $\pi/2$. Unentangled (product) states $\bigotimes_q |\psi_q\rangle$ form the manifold $\mathbb{M}_{\text{unentangled}} = (\mathbb{CP}^1)^N$. Their overlaps factorize, $\langle\psi|\phi\rangle = \prod_q \langle\psi_q|\phi_q\rangle$ (Lemma 4), so their fidelity can be computed classically in $\mathcal{O}(N)$ time.

**Encodings and circuits.** Data enters through a feature map $x \mapsto |\psi(x)\rangle$ built from gates. Every single-qubit gate is a Bloch rotation $\exp(-i\theta P/2)$ about a Pauli axis $P$, so two rotations about different axes place a qubit anywhere on $\mathbb{CP}^1$ using two real parameters per qubit, the dense angle encoding of LaRose & Coyle (2020) (Appendix B.2). Local rotations act on each factor independently and never leave $\mathbb{M}_{\text{unentangled}}$; reaching the rest of $\mathbb{CP}^{2^N-1}$ requires entangling gates. A parameterized quantum circuit (PQC) alternates rotation layers with entangling blocks, $U(\theta) = \prod_l W_l V_l(\theta_l)$, and hybrid QMeL models either train $\theta$ after a fixed encoding or re-upload the data between layers (Pérez-Salinas et al., 2020), which turns the model into a truncated Fourier series whose spectrum is set by the encoding schedule rather than by the trained parameters (Schuld et al., 2021).

**A.3 Reachable states and the dynamical Lie algebra.** A parameterized quantum circuit (PQC) applies gates $\exp(-i\theta_l H_l)$ with Hermitian generators $H_l$. Gates with non-commuting generators do not compose additively, so a deep circuit can also move the state along nested commutators of the generators. Their real span is the dynamical Lie algebra (DLA),

$$\mathfrak{g} = \mathrm{span}_{\mathbb{R}}\big\{\, iH_l,\ [iH_l, iH_m],\ \big[iH_l, [iH_m, iH_n]\big],\ \ldots \big\} \subseteq \mathfrak{su}(2^N), \tag{A.1}$$

so $\dim \mathfrak{g} \leq 4^N - 1$ (Appendix B.4). The states a circuit can reach form the orbit of $|0\rangle^{\otimes N}$ under $e^{\mathfrak{g}}$, a manifold of dimension at most $\dim \mathfrak{g}$. Single-qubit rotations give $\mathfrak{g} = \bigoplus_q \mathfrak{su}(2)$, of dimension $3N$, whose orbit is exactly $\mathbb{M}_{\text{unentangled}}$; generic entangling layers give $\mathfrak{su}(2^N)$, of dimension $4^N - 1$, whose orbit is all of $\mathbb{CP}^{2^N-1}$.

**A.4 Barren plateaus and kernel concentration.** A cost $C(\theta) = \langle\psi(\theta)|O|\psi(\theta)\rangle$ has a *barren plateau* when $\mathbb{E}_\theta[\partial_\mu C] = 0$ and $\mathrm{Var}_\theta[\partial_\mu C] \in \mathcal{O}(b^{-N})$ for some $b > 1$: by Chebyshev's inequality the gradient is then exponentially small with high probability and cannot be resolved with polynomially many shots (McClean et al., 2018; Cerezo et al., 2021b). When the generators, observable and input state lie in $\mathfrak{g}$, the variance scales as $1/\dim \mathfrak{g}$ times the $\mathfrak{g}$-purities of the input state and the observable (Ragone et al., 2024). For global fidelity losses ($O = |\phi\rangle\langle\phi|$), however, the observable has non-local components up to weight $N$, and Cerezo et al. (2021b) proved that global cost functions incur exponential barren plateaus ($\Theta(c^{-N})$) across both unentangled and entangled spaces (Lemma 2), while local observables ($k \ll N$) preserve polynomial gradients (Appendix B.4). *Kernel concentration* is a separate failure that depends on how inputs are spread over the state space and occurs even with no trainable parameters (Thanasilp et al., 2024). For Haar-random states $\mathbb{E}[F] = 2^{-N}$ and $\mathrm{Var}[F] = \Theta(4^{-N})$, so the Gram matrix $K_{ij} = F(x_i, x_j)$ approaches the identity, the loss cannot tell which pairs are more similar, and resolving its entries on hardware takes exponentially many shots (Miroszewski et al., 2024; Wang et al., 2021).

# B Extended Background

This appendix expands Section A. Section B.1 states the metric-learning objectives referred to in Section A.1, Section B.2 fixes the quantum notation used throughout, Section B.3 shows why the hybrid models of Section 2.1 are kernel methods, and Section B.4 collects the trainability and concentration results that Section 2.2 builds on.

## B.1 Metric-Learning Objectives

The objectives below are those used in this work and in the quantum metric-learning literature. They differ in how negatives are chosen and in what represents a class, but each defines a margin in $\mathcal{M}$ and each rewards an encoder $f_w$ that pushes negatives apart. The second property is the one the analysis turns on: in a space where nearly all pairs are already far apart, an objective can be satisfied without learning a geometry that transfers.

**Triplet loss.** Given an anchor $x_a$, a positive $x_p$ from the same class and a negative $x_n$ from a different class, the triplet loss (Weinberger et al., 2005; Schroff et al., 2015; Hoffer & Ailon, 2018)

$$\mathcal{L}_{\text{triplet}} = \max\{0,\ d\big(f_w(x_a), f_w(x_p)\big) - d\big(f_w(x_a), f_w(x_n)\big) + m\} \tag{B.1}$$

enforces a separation of at least $m > 0$ between positive and negative pairs. The hinge is one-sided: once a triplet satisfies the margin it contributes no gradient, however its negative is placed beyond that point.

**Prototypical loss.** In an episode with $K$ classes, the prototype of class $k$ is the mean embedding of its $n_k$ support examples, $\mathbf{c}_k = \frac{1}{n_k}\sum_{i:y_i=k} f_w(x_i)$, and each query $x_q$ is scored against every prototype (Snell et al., 2017),

$$\mathcal{L}_{\text{proto}} = -\log \frac{\exp\big(-d(f_w(x_q), \mathbf{c}_{y_q})\big)}{\sum_{k=1}^{K}\exp\big(-d(f_w(x_q), \mathbf{c}_k)\big)}, \tag{B.2}$$

with $d$ typically the squared Euclidean distance. The negatives are the $K-1$ competing prototypes, so their number is fixed by the episode rather than the batch. The prototype is an arithmetic mean, which presumes that $\mathcal{M}$ is closed under averaging. The quantum state space is not: an average of state vectors is in general neither normalized nor invariant to their global phases. Appendix C.5 gives the fidelity-optimal replacement.

**Contrastive (InfoNCE) loss.** For a batch of $B$ paired samples $\{(x_i, y_i)\}$, InfoNCE (van den Oord et al., 2018), the objective of CLIP-style multi-modal training (Radford et al., 2021), is

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log \frac{\exp\big(s(f_w(x_i), g_w(y_i))/\tau\big)}{\sum_{j=1}^{B}\exp\big(s(f_w(x_i), g_w(y_j))/\tau\big)}, \tag{B.3}$$

with encoders $f_w$ and $g_w$ (shared in the uni-modal case), a bounded similarity $s$ and a temperature $\tau$. It differs from the hinge in two ways that matter here. It has no margin at which the gradient vanishes, so it keeps rewarding a wider gap between the positive and the $B-1$ in-batch negatives; and its number of negatives grows with the batch, so it is dominated by negative pairs.

**Proxy and angular-margin losses.** A third family learns a class representative jointly with the encoder. Center loss adds a pull term $\frac{1}{2}\sum_i \|f_w(x_i) - \mathbf{c}_{y_i}\|^2$ toward a learned class center (Wen et al., 2016), and ArcFace normalizes embeddings and class weights onto the hypersphere and adds a margin $m$ to the angle $\vartheta_{y_i}$ between an embedding and its class direction (Deng et al., 2019),

$$\mathcal{L}_{\text{ArcFace}} = -\log \frac{\exp\big(\kappa\cos(\vartheta_{y_i} + m)\big)}{\exp\big(\kappa\cos(\vartheta_{y_i} + m)\big) + \sum_{k\neq y_i}\exp\big(\kappa\cos\vartheta_k\big)}, \tag{B.4}$$

with scale $\kappa$. Both require a class representative and an angle that behaves additively, neither of which is automatic when the target space is a product of projective spaces and the similarity a product of per-factor overlaps.

### B.2 Quantum States, Gates and Encodings

**State space.** An $N$-qubit register has state space $(\mathbb{C}^2)^{\otimes N}$. States are unit vectors defined up to a global phase, so pure states live on $\mathbb{CP}^{2^N-1}$, a compact manifold of real dimension $2(2^N-1)$ carrying the fidelity and Fubini–Study distance of Eq. 1. Fidelity is invariant under any unitary applied to both arguments, $F(V|\psi\rangle, V|\phi\rangle) = F(|\psi\rangle, |\phi\rangle)$, which is the identity behind Eq. 2.

**The single-qubit and product manifolds.** For $N=1$, $\mathbb{CP}^1$ is identified with the Bloch sphere $S^2$: a state is a real unit vector $\mathbf{r}\in\mathbb{R}^3$, its fidelity with another is $F=\frac{1}{2}(1+\mathbf{r}\cdot\mathbf{r}')$, and $d_{\mathrm{FS}}$ is half the angle between the two Bloch vectors, so $\mathbb{CP}^1$ is isometric to a sphere of radius $\frac{1}{2}$, so a single qubit behaves like a normalized classical embedding. States that factorize as $|\psi\rangle = \bigotimes_{q=1}^{N}|\psi_q\rangle$ carry no entanglement and occupy the product manifold $\mathbb{M}_{\text{unentangled}} \cong (\mathbb{CP}^1)^N$, of real dimension $2N$ rather than $2(2^N-1)$. On $\mathbb{M}_{\text{unentangled}}$ the fidelity is a product of single-qubit overlaps (Lemma 4), computable classically in $\mathcal{O}(N)$.

**Definition 3** (Fubini–Study ball)**.** *Let $\mathcal{M}$ be $\mathbb{CP}^{2^N-1}$ or $\mathbb{M}_{\text{unentangled}}$, and let $\mu$ be its uniform probability measure: the unitarily invariant (Haar) measure on $\mathbb{CP}^{2^N-1}$, and its $N$-fold product over qubits on $\mathbb{M}_{\text{unentangled}}$. The Fubini–Study ball of radius $t\in[0,\frac{\pi}{2}]$ about $\psi\in\mathcal{M}$ is*

$$\mathcal{B}_t(\psi) \;=\; \big\{\phi\in\mathcal{M} : d_{\mathrm{FS}}(\psi,\phi)\le t\big\} \;=\; \big\{\phi\in\mathcal{M} : F(\psi,\phi)\ge \cos^2 t\big\},$$

*and the fraction of $\mathcal{M}$ it occupies is $\mu\big(\mathcal{B}_t(\psi)\big)$, the probability that a uniformly random state lies within distance $t$ of $\psi$. Invariance of $\mu$ makes this fraction independent of $\psi$, so we write $\mu(\mathcal{B}_t)$.*

**Gates as rotations.** Up to global phase, every single-qubit unitary is a rotation of the Bloch sphere: $\exp(-i\vartheta P/2)$ for a Pauli operator $P$ rotates the Bloch vector by $\vartheta$ about the $P$ axis. The two rotations used in this work are

$$R_Y(\theta)=\begin{pmatrix}\cos\frac{\theta}{2} & -\sin\frac{\theta}{2}\\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2}\end{pmatrix}, \qquad R_Z(\gamma)=\begin{pmatrix}e^{-i\gamma/2} & 0\\ 0 & e^{i\gamma/2}\end{pmatrix}, \tag{B.5}$$

and applied in sequence to $|0\rangle$ they prepare

$$R_Z(\gamma)R_Y(2\theta)|0\rangle = e^{-i\gamma/2}\big(\cos\theta\,|0\rangle + e^{i\gamma}\sin\theta\,|1\rangle\big), \tag{B.6}$$

which, dropping the global phase, is the polar form of a qubit with polar angle $2\theta$ and azimuth $\gamma$. Two rotations about distinct axes therefore reach all of $\mathbb{CP}^1$ with two real parameters, and applied qubit-wise they prepare the product state of Eq. 11. Local rotations never leave $\mathbb{M}_{\text{unentangled}}$; reaching the rest of $\mathbb{CP}^{2^N-1}$ requires entangling gates such as CNOT, CZ or $ZZ(\alpha)$.

**Parameterized quantum circuits.** A parameterized quantum circuit (PQC) alternates layers of single-qubit rotations with fixed entangling blocks,

$$U(\vartheta)=\prod_{l=1}^{L} W_l\,V_l(\vartheta_l), \qquad V_l(\vartheta_l)=\bigotimes_{q=1}^{N}\exp\big(-i\,\vartheta_{l,q}P_{l,q}/2\big), \tag{B.7}$$

and prepares $U(\vartheta)|0\rangle^{\otimes N}$. In variational models $\vartheta$ is trained by gradient descent, with hardware gradients from the parameter-shift rule (Mitarai et al., 2018; McClean et al., 2016; Cerezo et al., 2021a). The layer count, the entangling pattern in $W_l$ and the generators $P_{l,q}$ together fix the reachable states, and with them the expressivity and trainability of Section B.4.

**Encoding classical data.** Angle encoding writes each real feature into the angle of one rotation, $|x\rangle = R_Y(x)|0\rangle = \cos(x/2)|0\rangle + \sin(x/2)|1\rangle$. A single rotation axis reaches only a great circle of each Bloch sphere, which is the access the encoders of Section 2.1.1 give the classical network. A second rotation about a different axis, as in Eq. B.6, gives the *dense* angle encoding of LaRose & Coyle (2020), with two real features per qubit and access to the full sphere; this is the parameterization QPMeL uses. Interleaving encoding and variational layers instead re-uploads the data (Pérez-Salinas et al., 2020), the family of Section 2.1.2.

**Estimating fidelity on hardware.** Fidelity is not an observable and must be estimated from repeated measurements. The *inversion* (compute–uncompute) test prepares $|\psi(x)\rangle$ on $N$ qubits, applies the inverse of the circuit preparing $|\psi(x')\rangle$, and estimates the probability of the all-zero outcome, which equals $F$. The *SWAP* test prepares the two states on separate registers, applies controlled-SWAPs from an ancilla, $2N+1$ qubits in all, and recovers $F = 2P(0) - 1$ from the ancilla alone. Both estimate a Bernoulli parameter from $S$ shots, so the standard error of $\hat{F}$ is $\mathcal{O}(S^{-1/2})$ independent of $N$, while the fidelities to be resolved shrink with $N$ (Section B.4). Appendix E.3 quantifies the resulting shot budget.

### B.3 Quantum Kernel Models

A feature map $x \mapsto |\psi(x)\rangle$ defines the kernel $k(x, x') = |\langle\psi(x)|\psi(x')\rangle|^2$, which is positive semi-definite by construction: the overlaps form the Gram matrix of unit vectors, and taking the squared modulus entrywise is the Schur product of that matrix with its conjugate, which preserves positive semi-definiteness (Schuld & Killoran, 2019; Schuld, 2021). Quantum kernel methods estimate this Gram matrix on a device and fit a classical convex model, such as an SVM, on top of it (Havlíček et al., 2019). The feature map need not be fixed: quantum embedding kernels tune the encoding by maximizing kernel-target alignment (Cristianini et al., 2001; Cortes et al., 2012; Hubregtsen et al., 2022), a metric-learning objective stated on the Gram matrix, and Lloyd et al. (2020) train an embedding to separate classes by fidelity.

**Variational models are kernel methods.** Expanding an observable in its eigenbasis, $O = \sum_k \lambda_k |v_k\rangle\langle v_k|$, the measured value is a fixed combination of fidelities against reference states,

$$\langle\psi(x)|O|\psi(x)\rangle = \sum_k \lambda_k \left|\langle v_k|\psi(x)\rangle\right|^2 = \sum_k \lambda_k\, k(v_k, x), \tag{B.8}$$

and a variational block $V$ before the measurement only replaces $O$ by $V^\dagger O V$, rotating the $|v_k\rangle$ without leaving this form. The encoding, not the variational block, therefore fixes the hypothesis class (Schuld, 2021), although the two families can differ in what they learn efficiently in practice (Jerbi et al., 2023). The parameter-predicting setup of Eq. 3 is a kernel of this kind whose feature map is learned by a classical network, the quantum analogue of a deep embedding kernel (Le & Xie, 2018).

### B.4 Trainability and Expressivity

**Expressivity and the dynamical Lie algebra.** The expressivity of an ansatz is usually measured by how uniformly its outputs cover the state space: an ensemble matching the first two Haar moments is an approximate 2-design (Harrow & Low, 2009), a regime deep hardware-efficient circuits reach quickly (Holmes et al., 2022), and capacity measures such as the effective dimension tell the same story (Abbas et al., 2021). The dynamical Lie algebra (DLA) measures the same quantity from the ansatz alone (Larocca et al., 2022; Ragone et al., 2024). Write the gates of Eq. B.7 as $\exp(-i\vartheta_l H_l)$ with Hermitian generators $H_l$. Two such gates compose into a third of the same form only when their generators commute; otherwise the Baker–Campbell–Hausdorff expansion

$$e^A e^B = \exp\Big(A + B + \tfrac{1}{2}[A,B] + \tfrac{1}{12}\big[A,[A,B]\big] - \tfrac{1}{12}\big[B,[A,B]\big] + \dots\Big) \tag{B.9}$$

shows that repeated layers open directions given by nested commutators of the generators. The DLA $\mathfrak{g} \subseteq \mathfrak{su}(2^N)$ of Eq. A.1 is the Lie closure of $\{iH_l\}$ and is computed by a fixed-point iteration: start from $B_0 = \{iH_1, \dots, iH_M\}$, form all commutators of pairs of current elements, expand each in the Pauli basis, keep those linearly independent of the span accumulated so far, and repeat until a pass adds nothing. The iteration terminates because $\dim \mathfrak{su}(2^N) = 4^N - 1$ bounds the span, and $\dim \mathfrak{g}$ is the size of the resulting basis.

The reachable unitaries form the connected group $e^{\mathfrak{g}}$, so the reachable states are the orbit of $|0\rangle^{\otimes N}$ under it, a manifold of dimension at most $\dim \mathfrak{g}$. Three cases bracket the range. Mutually commuting generators give an abelian $\mathfrak{g}$ of dimension at most $M$ whose orbit is a torus. Single-qubit rotations on every qubit give $\mathfrak{g} = \bigoplus_q \mathfrak{su}(2)$ of dimension $3N$, whose orbit is exactly $\mathbb{M}_{\text{unentangled}}$, each factor being the Bloch sphere. Generic entangling layers close the algebra to $\mathfrak{su}(2^N)$, whose orbit is all of $\mathbb{CP}^{2^N-1}$. Structured ansätze sit in between; free-fermion and transverse-field Ising generators close

into polynomially sized algebras and are the standard way to keep $\dim \mathfrak{g}$ tractable (Larocca et al., 2022).

**Barren plateaus.** For a cost $C(\vartheta) = \langle\psi(\vartheta)|O|\psi(\vartheta)\rangle$, a barren plateau means $\mathbb{E}_\vartheta[\partial_\mu C] = 0$ and $\mathrm{Var}_\vartheta[\partial_\mu C] \in \mathcal{O}(b^{-N})$ for some $b > 1$ (McClean et al., 2018). Chebyshev's inequality,

$$\Pr\big[\,|\partial_\mu C| \geq \delta\,\big] \leq \frac{\mathrm{Var}_\vartheta[\partial_\mu C]}{\delta^2}, \tag{B.10}$$

then confines the gradient to within $\mathcal{O}(b^{-N/2})$ of zero with high probability, and resolving a gradient of that size from finite shots costs a budget growing as $\mathcal{O}(b^N)$. Vanishing variance alone is not a plateau: a gradient concentrated tightly around a nonzero value is still trainable, so a plateau also requires the concentration point to be zero and the cost itself to concentrate. When the generators, observable, and input state lie in the dynamical Lie algebra $\mathfrak{g}$, the variance scales as $\mathrm{Var}_\vartheta[\partial_\mu C] \propto \mathcal{P}_\mathfrak{g}(\rho)\,\mathcal{P}_\mathfrak{g}(O)/\dim\mathfrak{g}$ for input state $\rho$, where the $\mathfrak{g}$-purity $\mathcal{P}_\mathfrak{g}$ is the squared norm of its projection onto $\mathfrak{g}$ (Ragone et al., 2024). On the full algebra $\mathfrak{su}(2^N)$, a global rank-one projector observable has purity $\Theta(1)$, yielding $\mathrm{Var} \in \Theta(4^{-N})$ matching Haar integration (Cerezo et al., 2021b). On restricted subalgebras such as $\bigoplus_q \mathfrak{su}(2)$, however, the global projector has components up to weight $N$ outside $\mathfrak{g}$, yielding purities that vanish exponentially ($\mathcal{P}_\mathfrak{g} = N2^{-N}$); as proved by Cerezo et al. (2021b), global fidelity costs suffer exponential barren plateaus ($\Theta(c^{-N})$) across both unentangled and entangled spaces (Lemma 2). For local observables acting on few-body subsystems ($k \ll N$), by contrast, polynomial trainability is preserved (Cerezo et al., 2021b). In a hybrid model the vanishing gradient sits between the loss and the classical encoder, so it degrades the trainability of the classical parameters as well.

**Fidelity concentration.** Exponential concentration is the statement that $\Pr[\,|k(x,x') - \mu| \geq \delta\,] \leq \epsilon$ with $\delta \in \mathcal{O}(b^{-N})$ for essentially all pairs of inputs (Thanasilp et al., 2024). It is indexed by the data rather than the parameters, so it persists where gradients are well behaved and survives when the circuit has no trainable parameters at all; expressive embeddings, entanglement, hardware noise and qubit count each drive it. For Haar-random states on $\mathbb{CP}^{2^N-1}$, $F$ is $\mathrm{Beta}(1, 2^N - 1)$ distributed (Lemma 3), so

$$\mathbb{E}[F] = 2^{-N}, \qquad \mathbb{E}[F^2] = \frac{2}{2^N(2^N+1)}, \qquad \mathrm{Var}[F] = \Theta(4^{-N}). \tag{B.11}$$

The Gram matrix $K_{ij} = k(x_i, x_j)$ then approaches the identity: its off-diagonal entries become nearly equal, so a loss cannot tell which pairs are more similar. A model can still fit its training set, because the diagonal stays at 1, but only by memorizing it.

The product manifold is not exempt. For two independent uniform points on a Bloch sphere $\mathbf{r}\cdot\mathbf{r}'$ is uniform on $[-1,1]$, so $F_q = \frac{1}{2}(1 + \mathbf{r}\cdot\mathbf{r}')$ is uniform on $[0,1]$, and the product $F = \prod_q F_q$ of $N$ independent factors has $\mathbb{E}[F] = 2^{-N}$ and $\mathbb{E}[F^2] = 3^{-N}$, hence

$$\mathbb{E}[F] = 2^{-N}, \qquad \mathrm{Var}[F] = 3^{-N} - 4^{-N} = \Theta(3^{-N}). \tag{B.12}$$

Uniformly spread product states therefore concentrate around the same value as Haar-random states, with a variance larger by a factor approaching $(4/3)^N$: removing entanglement changes how tightly the fidelities gather, not where they gather. On hardware the estimator noise does not shrink with $N$ while the separations to be resolved do, so distinguishing two kernel entries costs shots growing exponentially in $N$ (Miroszewski et al., 2024; Wang et al., 2021).

# C Extended Analysis

This appendix collects the arguments that Section 2 states in brief.

## C.1 The setup bounds the families it replaces

Every pure-state model considered here maps a fixed input to $|\psi(x)\rangle = U(z)\,|0\rangle^{\otimes N}$, so the only thing an ansatz contributes is its reachable set $\mathcal{R}(U) = \{U(z)\,|0\rangle^{\otimes N} : z \in \Theta\}$. When $f_w$ supplies $z$

per input, any embedding whose image lies in $\mathcal{R}(U)$ is representable up to the approximation error of the network, so two ansätze with the same reachable set are interchangeable here. The families of Sections 2.1.1 and 2.1.2 hold their trainable parameters fixed across inputs, so each realizes one particular map into the same set, and $f_w$ can output those parameter values itself; the converse fails, since their parameters cannot vary with the input. Results obtained in this setup therefore bound what either family can achieve on the same ansatz. The equivalence concerns representable embeddings only: the parameterization still determines the gradients that reach $f_w$.

### C.2 The packing argument

Place a Fubini–Study ball $\mathcal{B}_{\varepsilon/2}$ of radius $\varepsilon/2$ (Definition 3) around each state of an $\varepsilon$-separated set. The balls are disjoint, since overlapping balls would place their centres closer than $\varepsilon$, so the number of states is at most $1/\mu(\mathcal{B}_{\varepsilon/2})$, where $\mu$ is the fraction of the volume of $\mathcal{M}$ that a ball occupies. If the set is as large as possible, no further state can be added, so every point of $\mathcal{M}$ lies within $\varepsilon$ of some state in the set; the balls of radius $\varepsilon$ then cover $\mathcal{M}$, and the number of states is at least $1/\mu(\mathcal{B}_\varepsilon)$. On $\mathbb{CP}^{2^N-1}$ the ball fraction is known exactly (Appendix E.1),

$$\mu(\mathcal{B}_t) = (\sin t)^m, \qquad m = \dim_{\mathbb{R}} \mathcal{M}, \tag{C.1}$$

where $m$ is the number of real parameters of a state: $m = 2(2^N - 1)$ on $\mathbb{CP}^{2^N-1}$, and $m = 2N$ on $\mathbb{M}_{\text{unentangled}}$, a polar and an azimuthal angle per qubit. Eq. C.1 is the Euclidean law that the volume of an $m$-dimensional ball grows as the $m$-th power of its radius, with the arc length $t$ replaced by $\sin t = \sqrt{1-F}$, the length of the component of one state orthogonal to the other. For small $t$, $\sin t \approx t$; at $t = \pi/2$ the ball is the whole space and the count degenerates. On the product manifold the same law holds only up to a bounded correction, which Lemma 1 carries as $\Delta$.

Both bounds follow by substituting Eq. C.1 into the two counts above, and both have the same form: the log-capacity of a manifold is its dimension times a constant set by the resolution. At $\varepsilon = \pi/2$, Eq. 5 degenerates and the capacity is the orthogonal count $2^N$ of Eq. 4. At $\varepsilon = \pi/4$, where two states still have fidelity $\frac{1}{2}$, $1/\sin\varepsilon = \sqrt{2}$ and the lower bound becomes $m \log_2 \sqrt{2} = m/2$, so $M(\mathcal{M}, \pi/4) \geq 2^{m/2}$: each of the $m/2$ complex dimensions doubles the number of states the bound guarantees. The full space has $2^N - 1$ complex dimensions, so $M(\mathbb{CP}^{2^N-1}, \pi/4) \geq 2^{2^N-1}$, with a matching upper bound of $2^{2.77(2^N-1)}$. Halving the separation from $\pi/2$ to $\pi/4$ thus raises the capacity of the full space from $2^N$ to exponential in $2^N$.

The product manifold grows far more slowly. A single qubit has six states pairwise at least $\pi/4$ apart, the octahedron $\{|0\rangle, |1\rangle, |\pm\rangle, |\pm i\rangle\}$, so $M(\mathbb{M}_{\text{unentangled}}, \pi/4) \geq 6^N$, while Eq. 5 caps it at $(4 + 2\sqrt{2})^N N^N$. At $N = 10$ the product manifold holds at most $10^{19}$ such states, where the full space holds at least $10^{307}$, and each further halving of $\varepsilon$ multiplies both counts by roughly $2^m$. The proof is given in Appendix E.1.

### C.3 Closure of the dynamical Lie algebra

Reaching more states requires a larger DLA $\mathfrak{g}$ (Section 2.2.1). When the observable or input state lies in the circuit's algebra, gradient variance scales inversely with the DLA dimension as $1/\dim \mathfrak{g}$ (Ragone et al., 2024); this condition fails for global fidelity on $\mathbb{M}_{\text{unentangled}}$ (Appendix E.1). For global fidelity observables, however, gradients collapse exponentially across both unentangled and entangled spaces (Lemma 2), and reaching all of $\mathbb{CP}^{2^N-1}$ with $\mathfrak{g} = \mathfrak{su}(2^N)$ incurs $\Theta(4^{-N})$ variance (Cerezo et al., 2021b).

For $\mathfrak{su}(2^N)$, Eq. 6 is quadratic in the dimension reached rather than proportional to it. The two terms split the algebra at a given state: $m$ directions move the state and the remaining $(m/2)^2$ leave it unchanged (they form its stabilizer), so all but $m$ of the $4^N - 1$ directions are invisible to the embedding.

Generic entanglement reaches $\mathfrak{su}(2^N)$ because an algebra is closed under commutators. In the Pauli basis, $\mathfrak{su}(2^N)$ has one generator per non-identity Pauli string, $4^N - 1$ of them, and local rotations occupy only the $3N$ strings of weight one (those acting on a single qubit). The commutator of two anticommuting Pauli strings is a third, so two-qubit couplings across a connected set of pairs generate

strings of weight two, these generate weight three together with the local rotations, and the closure continues until every string is included. At $N = 10$ local rotations provide 30 of the $1\,048\,575$ generators, and adding a generic coupling generates all of the rest.

Structured generator sets are the exception: they close on subalgebras of polynomial dimension (Larocca et al., 2022), and a model that reaches beyond $\mathbb{M}_{\text{unentangled}}$ at polynomial gradient cost must use one. These subalgebras form a discrete family rather than a continuum, and each fixes its own reachable submanifold, committing the embedding in advance to the states that submanifold contains. Elsewhere in QML, such a restriction is justified when a good solution is known to lie in the retained subspace; in metric learning, the size of the reachable set is the reason for using the quantum state space, so that justification is unavailable.

### C.4 Pairwise cost and locality of the product kernel

Classically, $n$ inputs are embedded once and all $n^2$ similarities follow from inner products of the stored vectors. A generic quantum state has no efficient classical description, so each pair needs its own circuit execution: $\mathcal{O}(n^2)$ circuits per batch at $S$ shots each. A CLIP-style batch of $10^4$ contains $5 \times 10^7$ pairs, or $5 \times 10^{10}$ circuit executions per step at $10^3$ shots. Retrieval is worse: scoring a query against a database costs $\mathcal{O}(|\mathrm{DB}|)$ circuit runs with no index to prune them, whereas classical search is sublinear (Malkov & Yashunin, 2020). Both costs revert to classical ones when the state is classically tractable, since on $\mathbb{M}_{\text{unentangled}}$ Lemma 4 returns $F$ in $\mathcal{O}(N)$ arithmetic.

For small misalignments the factorized fidelity is a Gaussian kernel. Writing $\beta_q$ for the Bloch angle between the two states of qubit $q$,

$$F = \prod_q \cos^2 \frac{\beta_q}{2} \;\approx\; \exp\Big(-\tfrac{1}{4} \textstyle\sum_q \beta_q^2\Big), \tag{C.2}$$

accurate to under a percent for misalignments near $0.1$ radians. Fidelity is thus a Gaussian (RBF) kernel whose bandwidth is set by the output scale of $f_w$ rather than chosen. Local kernel machines need a number of samples exponential in the input dimension to represent functions that vary on the kernel's own scale (Bengio et al., 2005), and the concentration of Section 2.1.3 drives the Gram matrix toward the identity, shrinking the effective bandwidth further. The effective degrees of freedom then equal the number of training points, and no non-trivial generalization bound survives. Shattering is therefore the standard failure of a kernel with too small a bandwidth, not a new quantum effect, and the remedy lies in the encoder rather than the space.

### C.5 Trivial minimizers and the choice of loss

A *trivial minimizer* is an arrangement of embeddings that attains the objective's minimum while encoding nothing about the data. Metric-learning losses are computed from the similarities inside tuples of embeddings, pairs for contrastive losses, triplets for the hinge, the whole batch for InfoNCE (Appendix B.1), so each reads a batch only through its Gram matrix $K_{ij} = k_w(x_i, x_j)$, with proxy losses adding the similarities to a set of class representatives. Setting $K = I$ places every negative pair at maximal distance and so minimizes any objective built from negatives whatever the samples were, and every orthonormal $n$-frame inside $\mathbb{M}_{\text{unentangled}}$ realizes it (Proposition 1), so such arrangements are abundant rather than rare.

**Saturating and non-saturating losses reach $K = I$ differently.** The triplet hinge contributes gradient only while a negative violates the margin, and under concentration the typical negative starts at $d_{\mathrm{FS}} \approx \pi/2$, leaving the term inactive before training begins. InfoNCE has no margin and its gradient vanishes only with the negative similarities themselves, so it pushes toward orthogonality for as long as any remains. This predicts which tasks shatter: class-level objectives on MNIST and Omniglot do not, whereas under InfoNCE a sufficiently expressive encoder does on THINGS-EEG2 (Section 4.3).

**A class representative is not a mean.** The representatives that prototypical, center and ArcFace losses require must themselves be states, and an arithmetic mean of state vectors is not one: it is neither normalized nor phase-invariant, since $|\psi\rangle$ and $e^{i\alpha}|\psi\rangle$ are the same state and average differently. The fidelity-optimal prototype maximizes $\frac{1}{n}\sum_i |\langle\phi|\psi_i\rangle|^2 = \langle\phi|\,\bar{\rho}\,|\phi\rangle$ for $\bar{\rho} = \frac{1}{n}\sum_i |\psi_i\rangle\langle\psi_i|$, and

is its top eigenvector; restricted to $\mathbb{M}_{\text{unentangled}}$ the same maximization is a best rank-one tensor approximation, with no closed form and NP-hard in general. The additive surrogate $\sum_q F_q$ separates over qubits and is maximized by the top eigenvector of each reduced state $\bar{\rho}_q$, the per-qubit Bloch mean, a second reason to prefer it to the product kernel (Section 3.3).

### C.6 Smoothness

Smoothness requires the composite map $x \mapsto |\psi_w(x)\rangle$ to respond proportionally to its input. A small change in the encoder's output must move $d_{\text{FS}}$ by a small amount, so that gradients remain informative, and inputs close in $\mathcal{X}$ must stay close in $\mathcal{M}$ while distinct inputs stay distinct, so that a margin won in training transfers to unseen points. The geometry enforces neither condition.

**The parameterization is periodic.** Gate angles enter as $\exp(-izP/2)$, and the state is unchanged (up to a global phase) by $z \mapsto z + 2\pi$, so $z \mapsto U(z)\,|0\rangle^{\otimes N}$ is only locally injective. An encoder with unbounded outputs therefore aliases: inputs arbitrarily far apart in $\mathcal{X}$ can be assigned the same state, and the objective reads that pair at fidelity one, as a positive.

**The Lipschitz constant is unconstrained.** Let $\kappa$ be the smallest constant with $d_{\text{FS}}(\psi_w(x), \psi_w(x')) \leq \kappa\, d_{\mathcal{X}}(x, x')$ for all $x, x'$. The geometry bounds the left side by $\pi/2$ but places no bound on $\kappa$, since $f_w$ may cross that diameter over an arbitrarily small input distance and send neighbouring inputs to orthogonal states. A margin won in training therefore certifies nothing between the training points: the embedding can satisfy every constraint on $\mathcal{D}$ while varying arbitrarily fast elsewhere, which is geometric shattering (Proposition 1) stated as a property of the map. Generalization bounds for quantum embeddings scale with $\kappa$ rather than with the dimension of the Hilbert space (Kim & Bekiranov, 2022), so $\kappa$ is the quantity a design must control. Bounding it also removes the aliasing above, by capping the angle a bounded input range can sweep.

**Shattering is an encoder-side failure.** It needs $n$ mutually orthogonal reachable states, and removing entanglement does not remove them: the $2^N$ basis states of Eq. 4 lie inside $\mathbb{M}_{\text{unentangled}}$ and local rotations reach every one. A severe enough restriction does remove them, since a permutation-symmetric circuit never leaves the symmetric subspace, which holds only $N + 1$ mutually orthogonal states. But this caps the orthogonal capacity at $N + 1$, an advance commitment of the kind discussed in Appendix C.3, rather than a remedy. For any ansatz whose orthogonal capacity exceeds $n$, the constraint must fall on $f_w$.

## D QPMeL Architecture and Training Flow

Figure D.1 shows the split encoder of Section 3.1, and Figure D.2 the classical training flow of Section 3.2.

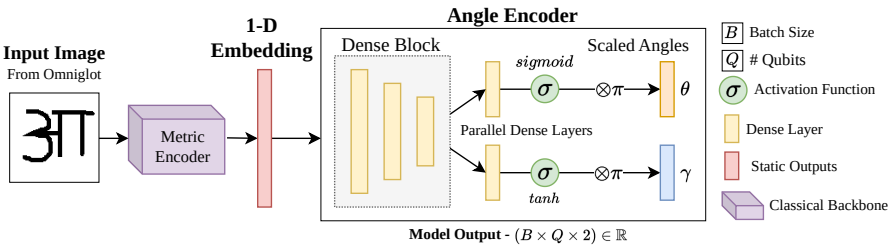


Figure D.1: *QPMeL architecture. A domain metric encoder maps the input to a latent vector, and an angle head projects it through bounded activations to $N$ coordinate pairs $(\theta_q, \gamma_q)$, one per qubit.*

## E Proofs and Derivations

Results are proved in the order the main text uses them: the capacity bounds of Section 2.2.1 (Appendix E.1), geometric shattering (Appendix E.2), the shot-noise scaling of Section 2.2.2 (Ap-

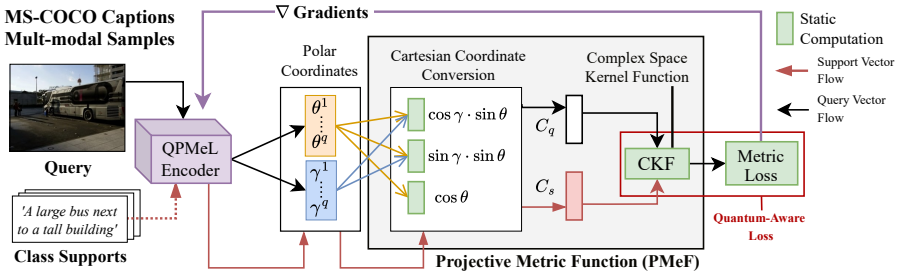


Figure D.2: *QPMeL training flow, shown for an MS-COCO image and caption. The PMeF (shaded) projects the encoder's angles onto real coordinate vectors $C_q$ and computes each qubit's fidelity with the CKF, and the metric loss on the resulting kernel sends gradients back to the encoder. Every block is classical, and no circuit runs during training.*

pendix E.3), and the closed-form product-state fidelity used in Sections 2.2.2 and 3 (Appendices E.4–E.6).

### E.1 Capacity of the Quantum State Space

This section proves Lemma 1 and Lemma 2. Throughout, $d = 2^N$, distances are Fubini–Study, and $\mu$ is the unitarily invariant probability measure on the state space, so $\mu(A)$ is the probability that a Haar-random state lies in $A$ and $\mu(\mathcal{B}_t)$ is the fraction of the space covered by a metric ball of radius $t$. Invariance is what the argument uses: every ball of a given radius carries the same measure, wherever its centre.

**Lemma 3** (Fubini–Study ball measure)**.** *For any fixed $|\psi\rangle \in \mathbb{CP}^{d-1}$ and $|\phi\rangle$ drawn from the unitarily invariant measure, $\Pr[F(\psi,\phi) \geq u] = (1-u)^{d-1}$ for $u \in [0,1]$. The measure of the ball $\mathcal{B}_t(\psi)$ of Definition 3 is therefore*

$$\mu\big(\mathcal{B}_t(\psi)\big) = \sin^{2(d-1)}(t), \qquad t \in [0, \tfrac{\pi}{2}]. \tag{E.1}$$

*Proof.* By unitary invariance take $|\psi\rangle = |0\rangle$, so $F = |\langle 0|\phi\rangle|^2$ is the first squared component of a uniform point on the unit sphere of $\mathbb{C}^d$, which is $\mathrm{Beta}(1, d-1)$ distributed. Its survival function is $\Pr[F \geq u] = (1-u)^{d-1}$. A state lies in $\mathcal{B}_t(\psi)$ exactly when $\arccos\sqrt{F} \leq t$, i.e. when $F \geq \cos^2 t$, and substituting $u = \cos^2 t$ gives $\mu(\mathcal{B}_t) = (1-\cos^2 t)^{d-1} = \sin^{2(d-1)}(t)$. □

*Proof of Lemma 1.* Let $\mathcal{S}$ be a maximal $\varepsilon$-separated subset of $\mathbb{CP}^{d-1}$, so $|\mathcal{S}| = M(\mathbb{CP}^{d-1}, \varepsilon)$. The balls $\{\mathcal{B}_{\varepsilon/2}(\psi)\}_{\psi\in\mathcal{S}}$ are pairwise disjoint, since two of them meeting would place their centres within $\varepsilon$ of each other, and they all sit inside a space of unit measure, so $|\mathcal{S}|\,\mu(\mathcal{B}_{\varepsilon/2}) \leq 1$. Maximality means no point can be added, so every point of the space lies within $\varepsilon$ of some centre and $\{\mathcal{B}_\varepsilon(\psi)\}_{\psi\in\mathcal{S}}$ covers it, giving $|\mathcal{S}|\,\mu(\mathcal{B}_\varepsilon) \geq 1$. Substituting Eq. E.1 with $d-1 = 2^N - 1$ and taking logarithms yields Eq. 5 with $m = 2(2^N-1)$ and $\Delta = 0$.

For the product manifold, $\mathbb{M}_{\text{unentangled}} \cong (\mathbb{CP}^1)^N$ carries the product metric and the product measure, and a ball of radius $\varepsilon$ about $\psi = \bigotimes_q \psi_q$ contains the product of the per-qubit balls $\mathcal{B}_{\varepsilon/\sqrt{N}}(\psi_q)$. Eq. E.1 at $d = 2$ gives $\mu(\mathcal{B}_t) = \sin^2 t$ on each factor, so

$$\sin^{2N}\!\left(\frac{\varepsilon}{\sqrt{N}}\right) \;\leq\; \mu\big(\mathcal{B}_\varepsilon(\psi)\big) \;\leq\; \sin^{2N}(\varepsilon),$$

the upper bound because the ball is contained in the product of the per-qubit balls of radius $\varepsilon$. The same two counting arguments then bracket $\log M(\mathbb{M}_{\text{unentangled}}, \varepsilon)$ between $2N\log(1/\sin\varepsilon)$ and $2N\log(\sqrt{N}/\sin(\varepsilon/2))$, which is Eq. 5 with $m = 2N$ and $\Delta = N\log N$, the slack coming from the $\sqrt{N}$ contraction of the per-qubit radius. □

*Proof of Lemma 2.* **Global fidelity barren plateau.** The global fidelity cost is $C_G(\theta) = 1 - F(\psi(\theta),\phi) = 1 - |\langle\psi(\theta)|\phi\rangle|^2$, which evaluates the global rank-1 projector $O_G = |\phi\rangle\langle\phi|$.

On $\mathbb{M}_{\text{unentangled}}$, the state factorizes into independent single-qubit states $|\psi(\theta)\rangle = \bigotimes_{q=1}^{N} |\psi_q(\theta_q)\rangle$, so the fidelity factorizes as $F = \prod_{q=1}^{N} F_q(\theta_q)$. For any parameter $\theta_1$ of qubit 1, the partial derivative is

$$\partial_{\theta_1} C_G = -\left(\frac{\partial F_1}{\partial \theta_1}\right) \prod_{q=2}^{N} F_q(\theta_q).$$

Under independent parameter distributions with $\mathbb{E}[\partial_{\theta_1} F_1] = 0$, the factors are mutually independent, so $\mathbb{E}[\partial_{\theta_1} C_G] = 0$ and the variance factorizes as

$$\mathrm{Var}_\theta[\partial_{\theta_1} C_G] = \mathbb{E}\Big[\big(\partial_{\theta_1} F_1\big)^2\Big] \prod_{q=2}^{N} \mathbb{E}\big[F_q^2\big].$$

If single-qubit states are Haar-distributed on each Bloch sphere $\mathbb{CP}^1$, $F_q \sim \mathrm{Beta}(1,1) = \mathrm{Uniform}(0,1)$, giving $\mathbb{E}[F_q^2] = 1/3$ and $\mathrm{Var}_\theta[\partial_{\theta_1} C_G] = \mathbb{E}[(\partial_{\theta_1} F_1)^2] \cdot 3^{-(N-1)}$. For single-qubit rotations $R_y(\theta_q)$ with uniform angles $\theta_q \in [-\pi, \pi]$ against $|0\rangle^{\otimes N}$ (the warm-up example of Cerezo et al. (2021b)), $F_q = \cos^2(\theta_q/2)$, giving $\mathbb{E}[F_q^2] = 3/8$, $\mathbb{E}[(\partial_{\theta_1} F_1)^2] = 1/8$ and $\mathrm{Var}_\theta[\partial_{\theta_1} C_G] = \frac{1}{8}(3/8)^{N-1}$. In the parameterization $R_y(2\theta_q)$ of Eq. 11, $F_q = \cos^2\theta_q$, and the prefactor becomes $\frac{1}{2}$ with the same base. Thus, on $\mathbb{M}_{\text{unentangled}}$, $\mathrm{Var}_\theta[\partial_\mu C_G] = \Theta(c^{-N})$ with $c = 1/\mathbb{E}[F_q^2] > 1$ set by the angle distribution and the target ($c = 3$ for Haar-distributed single-qubit states, $c = 8/3$ above), vanishing exponentially with $N$ despite zero entanglement. The bound concerns uniformly random parameters: a narrow initialization, or a trained encoder whose positives are aligned, has $\mathbb{E}[F_q^2] \to 1$ and escapes it. Moreover, by Proposition 1 of Cerezo et al. (2021b), for $\delta \in (0,1)$,

$$\Pr\big[C_G \le \delta\big] = \Pr\left[\prod_{q=1}^{N} \cos^2\frac{\theta_q}{2} \ge 1-\delta\right] \le (1-\delta)^{-1}\left(\frac{1}{2}\right)^N,$$

so the volume of parameter space within distance $\delta$ of the global minimum shrinks exponentially, forming a narrow gorge.

Note that the Lie-algebraic variance formula of Ragone et al. (2024), $\mathrm{Var}_\theta[\partial_\mu C] \propto \sum_j \mathcal{P}_{\mathfrak{g}_j}(\rho)\mathcal{P}_{\mathfrak{g}_j}(O)/\dim \mathfrak{g}_j$, requires $\rho \in i\mathfrak{g}$ or $O \in i\mathfrak{g}$. For $\mathfrak{g} = \bigoplus_q \mathfrak{su}(2)$, $i\mathfrak{g}$ contains only 1-body Pauli operators. The rank-1 state projector $\rho = |0\rangle\langle 0|^{\otimes N} = 2^{-N}\sum_S Z_S$ and observable $O_G = |\phi\rangle\langle\phi|$ have components up to weight $N$, so neither lies in $i\mathfrak{g}$, and their formal purities are $\mathcal{P}_{\mathfrak{g}}(\rho) = \mathcal{P}_{\mathfrak{g}}(O_G) = N \cdot 2^{-N}$, which vanish exponentially rather than remaining $\Theta(1)$.

On $\mathbb{CP}^{2^N-1}$, generic entangling circuits close to $\mathfrak{su}(2^N)$ with $\dim \mathfrak{g} = 4^N - 1$. When the differentiated gate sits between two independent unitary 2-designs on $\mathrm{SU}(2^N)$, both the traceless parts of $\rho$ and $O_G$ lie in $i\mathfrak{su}(2^N)$ with purities $\mathcal{P}_{\mathfrak{g}} = 1 - 2^{-N} = \Theta(1)$. Exact Haar integration over $\mathrm{SU}(2^N)$ yields $\mathrm{Var}_\theta[\partial_\mu C_G] = \frac{1}{2(2^N+1)^2} = \Theta(4^{-N})$ for Pauli generators (McClean et al., 2018; Cerezo et al., 2021b).

**Measurement dimension and local observable trainability.** By Theorem 2 and Corollary 2 of Cerezo et al. (2021b), when the observable $O$ is an $m$-local operator acting on $k$-qubit subsystems with $k \ll N$, the gradient variance for shallow circuits is lower-bounded by $\Omega(1/\mathrm{poly}(N))$. For the additive kernel $\bar{k}_w(x,x') = \frac{1}{N}\sum_{q=1}^{N} F_q(x,x')$ introduced in Section 3.3, the corresponding observable $O_L = \frac{1}{N}\sum_{q=1}^{N} |\psi_q(x')\rangle\langle\psi_q(x')| \otimes \mathbb{I}_{\bar{q}}$ is strictly 1-local ($k = 1$). Differentiating yields $\partial\bar{k}_w/\partial\theta_q = \frac{1}{N}\partial F_q/\partial\theta_q$, so

$$\mathrm{Var}_\theta\left[\frac{\partial \bar{k}_w}{\partial \theta_q}\right] = \frac{1}{N^2}\mathrm{Var}_\theta\left[\frac{\partial F_q}{\partial \theta_q}\right] = \Theta\left(\frac{1}{N^2}\right),$$

which scales polynomially with $N$ and avoids the narrow gorge, confirming the trainability of the local observable regime. □

**Tightness.** The volumetric lower bound is not tight on $\mathbb{M}_{\text{unentangled}}$. The six octahedral states $\{|0\rangle, |1\rangle, |\pm\rangle, |\pm i\rangle\}$ of a single qubit are pairwise at fidelity at most $\frac{1}{2}$, so any two distinct elements of

their $N$-fold product differ on some qubit and have $F = \prod_q F_q \le \frac{1}{2}$, giving $M(\mathbb{M}_{\text{unentangled}}, \pi/4) \ge 6^N$ where the volumetric bound returns $2^N$. Products of separated sets remain separated under a product fidelity, so constructions of this kind improve the constant but leave the $\Theta(m)$ scaling of $\log M$, which is all Lemma 1 is used for, unchanged.

### E.2 Geometric Shattering

*Proof of Proposition 1.* (i) By Lemma 4, $F = \prod_q F_q$ with each $F_q \in [0, 1]$. A product of factors in $[0, 1]$ is zero if and only if some factor is zero, and one if and only if every factor is one.

(ii) Write $s_{ij} = k_w(x_i, y_j) \in [0, 1]$. The $i$-th term of Eq. B.3 is

$$\ell_i = \log\Big(1 + \sum_{j \ne i} \exp\big((s_{ij} - s_{ii})/\tau\big)\Big),$$

which depends on the embeddings only through the gaps $s_{ij} - s_{ii}$. Each gap is at least $-1$, so $\ell_i \ge \log\big(1 + (B-1)e^{-1/\tau}\big)$, with equality if and only if $s_{ii} = 1$ and $s_{ij} = 0$ for every $j \ne i$. Averaging over $i$ gives the bound, attained exactly when $s_{ij} = \delta_{ij}$; the same argument applies to the columns of a symmetric loss. Now $s_{ii} = 1$ means $f_w(x_i)$ and $g_w(y_i)$ are the same state $\psi_i$, and $s_{ij} = 0$ for $i \ne j$ then means $\psi_1, \dots, \psi_B$ are mutually orthogonal. A set of $B$ mutually orthogonal states exists if and only if $B \le 2^N$ (Eq. 4), and for $B \le 2^N$ distinct computational basis states, which lie in $\mathbb{M}_{\text{unentangled}}$, realize it. The loss sees only the Gram matrix, which equals the identity for any assignment of samples to orthonormal states, so every such assignment attains the bound. When the training set itself has $n \le 2^N$ pairs, one assignment attains the bound on every batch simultaneously.

On the unit sphere $S^{d-1}$ with cosine similarity the per-pair floor is $-1$, which holds only for antipodal points. No three points are pairwise antipodal, so for $B > 2$ no batch reaches the floor, and the most even spread of $B$ unit vectors, a regular simplex with pairwise cosine $-1/(B-1)$, needs $d \ge B - 1$. □

### E.3 Fidelity Signal-to-Noise Ratio

This section derives the FSNR of Eq. 8 for the two estimators of Appendix B.2. Each estimates $F$ from $S$ shots with a conditional variance $\mathrm{Var}(\hat{F} \mid F)$ fixed by the estimator, and the FSNR follows by substituting the moments of $F$ from Eq. B.11 for Haar-random states and Eq. B.12 for uniformly spread product states.

**Inversion test.** The all-zero outcome is Bernoulli with parameter $F$, so $\mathrm{Var}(\hat{F} \mid F) = F(1-F)/S$ and

$$\mathrm{FSNR}_{\text{inv}} = S\,\frac{\mathbb{E}[F^2] - (\mathbb{E}[F])^2}{\mathbb{E}[F] - \mathbb{E}[F^2]}. \tag{E.2}$$

For Haar-random states the numerator is $(d-1)/\big(d^2(d+1)\big)$ and the denominator $(d-1)/\big(d(d+1)\big)$, while for uniformly spread product states on $\mathbb{M}_{\text{unentangled}}$ the moments are $\mathbb{E}[F] = 2^{-N}$ and $\mathbb{E}[F^2] = 3^{-N}$, so

$$\mathrm{FSNR}_{\text{inv}}^{\text{Haar}} = S\,2^{-N}, \qquad \mathrm{FSNR}_{\text{inv}}^{\text{prod}} = S\,\frac{3^{-N} - 4^{-N}}{2^{-N} - 3^{-N}} = \Theta\big(S\,(2/3)^N\big). \tag{E.3}$$

Holding the ratio constant therefore costs $S \in \Omega(2^N)$ under Haar-random entanglement and $S \in \Omega\big((3/2)^N\big)$ on the product manifold.

**SWAP test.** The ancilla returns 0 with probability $(1 + F)/2$ and $\hat{F} = 2\hat{P}(0) - 1$, so $\mathrm{Var}(\hat{F} \mid F) = (1 - F^2)/S$ and

$$\mathrm{FSNR}_{\text{SWAP}} = S\,\frac{\mathbb{E}[F^2] - (\mathbb{E}[F])^2}{1 - \mathbb{E}[F^2]}. \tag{E.4}$$

For Haar-random states the denominator is $(d-1)(d+2)/\big(d(d+1)\big)$, giving

$$\mathrm{FSNR}_{\text{SWAP}}^{\text{Haar}} = \frac{S}{2^N(2^N+2)} = \Theta\big(S\,4^{-N}\big), \tag{E.5}$$

so a constant ratio requires $S \in \Omega(4^N)$. These are ideal statistical rates; readout error, gate infidelity and crosstalk on hardware bias $\hat{F}$ and raise the effective cost further.

### E.4 Factorized Inner Product

**Lemma 4.** *For $N$-qubit product states $|\psi_a\rangle = \bigotimes_{i=1}^{N} |\psi_a^i\rangle$ and $|\psi_b\rangle = \bigotimes_{i=1}^{N} |\psi_b^i\rangle$,*

$$\langle\psi_a|\psi_b\rangle = \prod_{i=1}^{N} \langle\psi_a^i|\psi_b^i\rangle \,.$$

*Proof.* The adjoint of a tensor product is the tensor product of the adjoints in the same order, $\left(\bigotimes_i |\psi_a^i\rangle\right)^\dagger = \bigotimes_i \langle\psi_a^i|$; unlike a matrix product, the factors are not reversed. The mixed-product property $(A_1 \otimes \cdots \otimes A_N)(B_1 \otimes \cdots \otimes B_N) = A_1B_1 \otimes \cdots \otimes A_NB_N$ then gives

$$\langle\psi_a|\psi_b\rangle = \Big(\bigotimes_{i=1}^{N} \langle\psi_a^i|\Big)\Big(\bigotimes_{i=1}^{N} |\psi_b^i\rangle\Big) = \bigotimes_{i=1}^{N} \langle\psi_a^i|\psi_b^i\rangle = \prod_{i=1}^{N} \langle\psi_a^i|\psi_b^i\rangle \,,$$

since a tensor product of scalars is their ordinary product. □

### E.5 Complex Kernel Function (CKF)

**Lemma 5.** *Let angular encodings $a$ and $b$ prepare the single-qubit states $|\psi_a\rangle$ and $|\psi_b\rangle$, with real coordinate vectors $(x, y, z)$ as in Eq. 12. Then*

$$|\langle\psi_a|\psi_b\rangle|^2 = (z_a \cdot z_b + x_a \cdot x_b + y_a \cdot y_b)^2 + (x_a \cdot y_b - y_a \cdot x_b)^2$$

*Proof.* Based on the spherical coordinates we create, we can assume:

$$\begin{aligned} x &= \cos(\gamma) \cdot \sin(\theta) \\ y &= \sin(\gamma) \cdot \sin(\theta) \\ z &= \cos(\theta) \end{aligned}$$

As we use only $R_Y$ and $R_Z$ gates, our qubit states can be written as functions of the $x, y, z$ defined above after we expand the exponential term:

$$|\psi\rangle = z\,|0\rangle + (x + iy)\,|1\rangle$$

Their inner product is therefore

$$\begin{aligned} \langle\psi_a|\psi_b\rangle &= (z_a\,\langle 0| + (x_a - iy_a)\,\langle 1|) \cdot (z_b\,|0\rangle + (x_b + iy_b)\,|1\rangle) \\ &= z_a \cdot z_b + (x_a - iy_a) \cdot (x_b + iy_b) \\ &= z_a \cdot z_b + x_a \cdot x_b + ix_a \cdot y_b - iy_a \cdot x_b - i^2 y_a \cdot y_b \\ &= (z_a \cdot z_b + x_a \cdot x_b + y_a \cdot y_b) + i(x_a \cdot y_b - y_a \cdot x_b) \\ &= \lambda_r + i\lambda_c \end{aligned}$$

where $\lambda_r = z_a \cdot z_b + x_a \cdot x_b + y_a \cdot y_b$ and $\lambda_c = x_a \cdot y_b - y_a \cdot x_b$ are both real numbers computed using real inner products. Applying the magnitude square :

$$\begin{aligned} |\langle\psi_a|\psi_b\rangle|^2 &= (\lambda_r + i\lambda_c) \cdot (\lambda_r - i\lambda_c) = \lambda_r^2 - i^2\lambda_c^2 = \lambda_r^2 + \lambda_c^2 \\ &= (z_a \cdot z_b + x_a \cdot x_b + y_a \cdot y_b)^2 + (x_a \cdot y_b - y_a \cdot x_b)^2 \end{aligned}$$

Hence proved. □

### E.6 Projective Metric Function Formula

**Lemma 6.** *Given any pair of angular encodings $a, b$ such that they create the $N$-qubit product states $|\psi_a\rangle, |\psi_b\rangle$, their similarity in Hilbert Space can be computed via the single qubit CKF (see E.5) as follows:*

$$|\langle\psi_a|\psi_b\rangle|^2 = \prod_{i=1}^{N} \left((z_a^i \cdot z_b^i + x_a^i \cdot x_b^i + y_a^i \cdot y_b^i)^2 + (x_a^i \cdot y_b^i - y_a^i \cdot x_b^i)^2\right) = \prod_{i=1}^{N} \left((\lambda_r^i)^2 + (\lambda_c^i)^2\right)$$

*Proof.* We know from E.4 -

$$\langle \psi_a | \psi_b \rangle = \prod_{i=1}^{N} \langle \psi_a^i | \psi_b^i \rangle = \langle \psi_a^1 | \psi_b^1 \rangle \times \cdots \times \langle \psi_a^N | \psi_b^N \rangle$$

Applying the magnitude operator on both sides:

$$| \langle \psi_a | \psi_b \rangle | = | \langle \psi_a^1 | \psi_b^1 \rangle \times \cdots \times \langle \psi_a^N | \psi_b^N \rangle |$$

Notice that each inner product $\langle \psi_a^i | \psi_b^i \rangle$ can be replaced by $\lambda^i \in \mathbb{C}$.

$$| \langle \psi_a | \psi_b \rangle | = |\lambda^1 \times \cdots \times \lambda^N|$$

We can then apply the property that $|z_0 \cdot z_1| = |z_0| \cdot |z_1| \;\; \forall z_0, z_1 \in \mathbb{C}$ to get:

$$| \langle \psi_a | \psi_b \rangle | = |\lambda^1| \times \cdots \times |\lambda^N| = | \langle \psi_a^1 | \psi_b^1 \rangle | \times \cdots \times | \langle \psi_a^N | \psi_b^N \rangle | = \prod_{i=1}^{N} | \langle \psi_a^i | \psi_b^i \rangle |$$

Finally, as each magnitude of complex number yields a real number, we can apply the square operator (for real $a_i$, $(a_1 \cdots a_N)^2 = a_1^2 \cdots a_N^2$) to get:

$$| \langle \psi_a | \psi_b \rangle |^2 = \prod_{i=1}^{N} | \langle \psi_a^i | \psi_b^i \rangle |^2$$

Replacing the fidelities with the CKF function calculated earlier in section E.5:

$$\begin{aligned} | \langle \psi_a | \psi_b \rangle |^2 &= \prod_{i=1}^{N} \left( (z_a^i \cdot z_b^i + x_a^i \cdot x_b^i + y_a^i \cdot y_b^i)^2 + (x_a^i \cdot y_b^i - y_a^i \cdot x_b^i)^2 \right) \\ &= \prod_{i=1}^{N} \left( (\lambda_r^i)^2 + (\lambda_c^i)^2 \right) \end{aligned}$$

Hence proved. □

# F Additional Results

## F.1 Few-shot baselines

Baseline figures are those each paper reports; QPMeL is evaluated on the same classes, ways and shots, but inputs and encodings are not matched. Thumwanit et al. (2021) downsample images to $4 \times 4$ binary pixels and encode them through a trained lookup table of $U_3$ gates, and Wang et al. (2024) amplitude-encode $8 \times 8$ images. For Wang et al. (2024) Table 1 reports the strongest model in their table for each task: QDiff, or QuantumNAT on 3-way 1-shot Fashion-MNIST. QPMeL uses fewer qubits than the baseline in eight comparisons and the same number in two; the exception is Huang et al. (2025) on Omniglot (11 against 9). Liu et al. (2022) feed a CNN's outputs to the metric-learning ansatz of Lloyd et al. (2020), which re-uploads the encoding between trainable $ZZ$ and $R_Y$ layers, and classify by fidelity to class prototypes. Their ablation places an unentangled variant at about $97.8\%$ against $98.3\%$ for the weakly entangled model, and a wider initialization, which entangles more, at about $97.3\%$ (read from their Figures 2 and 3, which report the best test accuracy reached during training).

## F.2 THINGS-EEG2

All heads share the encoders, loss and optimizer of Chen et al. (2024): 10 qubits, AdamW at learning rate $2 \times 10^{-4}$, 200 epochs, batch size 1024 and a learned logit scale initialized at $1/0.07$. Table 2 reports subject 1 over five seeds, except one- and two-layer DR on QNervFormer (ten and three seeds). Its QNervFormer columns are at epoch 200. The QSTConv and branch-dropout columns select the checkpoint with the lowest loss on 740 held-out training pairs, never on the test set; on QSTConv this moves held-out accuracy by at most $0.4$ points from epoch 200 for the four-layer heads and leaves

QPMeL at $19.9\%$ ($25.7\%$ train at epoch 200). On QNervFormer, QPMeL ends at a training loss of $0.62$ and one-layer DR between $1.6$ and $4.7$, while the four-layer circuit heads end between $5.9$ and $6.9$, near $\ln 1024 = 6.93$, the loss of a uniform softmax over the batch. Under the $\langle Z \rangle$ cosine readout of Chen et al. (2024), DR fits less than under fidelity: $0.8 \pm 0.4\%$ of QNervFormer's training images and $0.5 \pm 0.6\%$ of QSTConv's at four layers, and at most $3.3\%$ at any depth. Our E-VQC matches the equations of Chen et al. (2024) to $1.8 \times 10^{-7}$ but reaches $0.84\%$ held-out on QSTConv against their $3.5\%$ average over subjects; the paper has no public code, leaves seven architectural choices unspecified and does not fully specify its evaluation protocol.

**Regularization.** The branch-dropout columns of Table 2 add dropout $0.5$ at the end of each convolution branch of QNervFormer, and Table F.3 applies the same protocol to DR at every depth. QPMeL's $11.8 \pm 5.6\%$ spread there comes from one seed that does not take off within 200 epochs; of the eight runs of this configuration, seven reach $12.0$–$17.5\%$ held-out (median $15.0\%$). Attention dropout, feed-forward dropout or per-sample normalization alone leaves memorization in place ($86$–$98\%$ train at epoch 200). Holding QPMeL's kernel additive for all of training does not prevent memorization (Table F.1), and fixing the logit scale at 5 slows it to $87.7\%$ train without raising held-out accuracy.

**Curriculum ablation.** Table F.1 holds QPMeL's kernel additive for all 200 epochs, so it never becomes the product PMeF, and compares it with the additive-to-product curriculum of Eq. 16 over the full grid of 10 subjects and 5 seeds. On QSTConv the product end of the curriculum adds about 5 points of top-1. On QNervFormer both schedules memorize: the additive kernel's mean unmatched fidelity over a batch of $B$ is at least $\frac{1}{2}\big(1 - 1/(B-1)\big)$, so it cannot place the batch at orthogonal states, yet its training loss still falls to $1.40$ with held-out accuracy at chance.

Table F.1: *QPMeL kernel schedule on THINGS-EEG2, 10 subjects and 5 seeds: 200-way held-out accuracy (%), final training loss, and runs whose final loss stays above* $6.9$ *of* $50$*. Here, Curr. is Curriculum and Add. is Additive.*

| | Top-1 | | Top-5 | | Final loss | | Collapsed | |
|---|---|---|---|---|---|---|---|---|
| Encoder | Curr. | Add. | Curr. | Add. | Curr. | Add. | Curr. | Add. |
| QSTConv | $\mathbf{23.96 \pm 5.2}$ | $19.11 \pm 3.9$ | $\mathbf{56.82 \pm 7.6}$ | $48.56 \pm 5.5$ | 4.81 | 5.38 | 0 | 0 |
| QNervFormer | $0.92 \pm 0.8$ | $0.73 \pm 0.5$ | $4.92 \pm 2.0$ | $3.93 \pm 1.4$ | 0.50 | 1.40 | 0 | 0 |

### F.3 Kernel and gradient diagnostics

Table F.2 collects the measurements behind Section 4.3 on QNervFormer: kernel statistics at initialization on 10 qubits with four circuit layers and inputs drawn uniformly from $[-\pi, \pi]$, the RMS gradient each head passes to the encoder at the first epoch, and the fit to training images after 200 epochs. Range is the spread of kernel values times CLIP's initial logit scale of $14.29$, the logit separation available to the softmax, and structure is the rank correlation $\rho$ of the kernel with input similarity; QPMeL's structure was not measured (n/m). Each failing kernel lacks one of the two. Under fidelity the E-VQC's trainable layers cancel (Eq. 2), leaving the single-axis $R_Y$ product kernel, structured ($\rho = 0.79$) but concentrated at $2^{-10}$ and worth $0.09$ nats. The $\langle Z \rangle$ cosine readout has range ($6.6$ nats) but little structure ($\rho = 0.19$). Concentration does not require entanglement: QPMeL's product kernel also starts at $2^{-10}$, and its additive kernel is what gives it $1.39$ nats. No failing head has a vanishing gradient. The $\langle Z \rangle$ E-VQC passes the encoder 17 times QPMeL's gradient at the first epoch and fits 30 times worse, and scoring the E-VQC by the log of its fidelity gives the largest gradient ($955$) and the worst fit ($0.3\%$). The E-VQC's circuit-gradient variance $\mathrm{Var}[\partial \langle Z_0 \rangle / \partial w]$ falls from $0.22$ to $2.3 \times 10^{-3}$ between 2 and 12 qubits, and its per-parameter gradient in mid-training shows no depth trend from 1 to 16 layers that survives a rerun. Redrawing the E-VQC's trainable weights changes its fidelity kernel by at most $1.6 \times 10^{-6}$ of its mean, as Eq. 2 requires. Its entangling layers shrink the single-qubit readout, with mean $|\langle Z_i \rangle|$ falling from $0.27$ at one layer to $0.10$ at four, and cosine renormalizes what remains.

**Re-uploading depth.** Table F.3 varies the number of layers $L$ of the two circuits scored by fidelity. Each DR layer applies a trainable $R_Y(w_q)$ and a data $R_Z(x_q)$ to every qubit, then a CNOT ring. The

Table F.2: *Kernel and gradient diagnostics on THINGS-EEG2 with QNervFormer; four-layer circuits.*

| | | Kernel at initialization | | | | |
|---|---|---|---|---|---|---|
| Head | Similarity | Mean | Range (nats) | Structure $\rho$ | Encoder gradient | Train (%) |
| E-VQC | $\langle Z \rangle$ cosine | $-3.9 \times 10^{-3}$ | 6.57 | 0.19 | 10.6 | 3.2 |
| E-VQC | Fidelity | $8.8 \times 10^{-4}$ | 0.09 | 0.79 | 0.010 | 10.1 |
| E-VQC | log fidelity | $8.8 \times 10^{-4}$ | – | 0.79 | 955 | 0.3 |
| DR | Fidelity | $1.0 \times 10^{-3}$ | 0.02 | 0.06 | 0.021 | 6.5 |
| QPMeL | PMeF, product | $9.9 \times 10^{-4}$ | 0.07 | n/m | 0.62 | **96.5** |
| | PMeF, additive | 0.50 | 1.39 | n/m | | |

ring after the last upload is the same unitary in both towers and cancels from the fidelity (Eq. 2), so the kernel contains $L - 1$ rings. At $L = 1$ it is the product kernel $\prod_q |\langle 0| R_Y(-w_q)\, R_Z(x'_q - x_q)\, R_Y(w'_q) |0\rangle|^2$ on $\mathbb{M}_{\text{unentangled}}$, with primes marking the image tower, in which the encoder sets only the azimuth of each qubit and the polar angle is a trained constant of each tower. When the two towers' angles agree, qubit $q$ contributes $1 - \sin^2 w_q \sin^2((x_q - x'_q)/2)$, so $w_q$ sets how strongly the qubit separates inputs, from not at all at $w_q = 0$ to the E-VQC's $\cos^2((x_q - x'_q)/2)$ at $w_q = \pi/2$. On the memorizing QNervFormer this is the best-fitting circuit head in the study ($65.0 \pm 17.9\%$ over ten seeds, range 40.0–85.5%), and at one and two layers DR fits more than the E-VQC, whose trainable layers cancel. Each added layer removes structure from DR's kernel, and fit falls with it; by $L = 8$ the kernel's mean and spread match those of Haar-random states. The E-VQC's kernel is the same at every depth within a tower, and its fit settles near 10% from four layers on. No run on this backbone exceeds 3.0% held-out.

On the two backbones that generalize, held-out accuracy falls with depth as fit does. One-layer DR reaches $12.3 \pm 5.2\%$ held-out on QSTConv and $6.6 \pm 4.6\%$ on QNervFormer with branch dropout, against $19.9 \pm 3.1\%$ and $11.8 \pm 5.6\%$ for QPMeL under the same selection; its best seed on QSTConv reaches 16.5%, but its mean trails QPMeL on both. Two layers, which leave one ring in the kernel, give 3.7% and 2.3%, and four or more stay within half a point of chance. The circuit head that generalizes best is therefore the one whose entanglement cancels, and it trails QPMeL, which differs from it both in setting both angles of every qubit and in beginning training with the additive kernel; these factors are not separated here.

The first-epoch encoder gradient rises with depth, so the collapse is not a vanishing gradient at initialization. The one-layer circuit passes the smallest gradient because its towers' polar angles start as independent draws: the kernel between towers then starts concentrated, averaging $3.6 \times 10^{-4}$ at initialization against $2^{-10} \approx 9.8 \times 10^{-4}$, and its gradient is multiplicative (Eq. 9). Every one-layer run starts on the $\ln 1024$ plateau, and when it leaves sets most of the spread. On QNervFormer runs leave between epochs 30 and 90, and the epoch of departure has rank correlation $-0.81$ with the final fit over ten seeds. With branch dropout they leave later, between epochs 75 and 125, and the one seed that never leaves ends at 0.5% held-out. On QSTConv all five seeds leave by epoch 100, but two climb more slowly and end at 5.5 and 8.0% against 15.5–16.5% for the rest. On these two backbones deeper circuits leave later or not at all: two-layer runs leave between epochs 75 and 150, and no four-layer run on QSTConv leaves. QPMeL's loss instead falls from the first epoch on both regularized backbones, as the range of its additive kernel allows (Table F.2). Circuit gradients measured at epoch 100 do fall with depth, by $40\times$ from one layer to sixteen, but by then the shallow runs have left the plateau and the deep ones have not. At four layers the separate five-seed runs of Table 2 give $6.5 \pm 1.7\%$ train.

**Trained kernels.** Table F.4 measures the kernel each run ends with, on the first 1024 training pairs at the checkpoint Table 2 reports. For Haar-random states on 10 qubits both kinds of pair average $2^{-10}$, and a fraction $1 - (1 - 0.01)^{1023} > 0.9999$ of unmatched pairs falls below 0.01. With depth, DR's kernel moves to these values on both backbones: the gap between matched and unmatched pairs closes and the fraction of unmatched pairs below 0.01 rises, reaching the Haar value at four layers on QSTConv and at sixteen with branch dropout. QPMeL ends with the largest gap of any generalizing run on both backbones, 0.29 on QSTConv against 0.12 for one-layer DR, with only 0.2% of unmatched pairs below 0.01. Unregularized QNervFormer under QPMeL instead sends

Table F.3: *Circuit depth under the fidelity readout on THINGS-EEG2. Structure is measured at initialization within one tower on inputs uniform in* $[-\pi, \pi]$*, and the encoder gradient is the RMS at the first epoch on QNervFormer. Train is the fit to 200 training images on QNervFormer after 200 epochs, three seeds (*$^{\ddagger}$*ten). Held-out rows are on QSTConv and on QNervFormer with branch dropout, with the checkpoint selection of Table 2, five seeds (*$^{\dagger}$*four); QPMeL reaches* 19.9 *and* 11.8% *there. n/r, not run.*

| | | Layers $L$ | | | | |
|---|---|---|---|---|---|---|
| Head | | 1 | 2 | 4 | 8 | 16 |
| DR | Rings in the kernel | 0 | 1 | 3 | 7 | 15 |
| | Structure $\rho$ | 0.64 | 0.46 | 0.06 | 0.00 | 0.00 |
| | Encoder gradient | $1.4 \times 10^{-4}$ | $8.3 \times 10^{-3}$ | 0.022 | 0.041 | 0.068 |
| | Train (%) | $\mathbf{65.0 \pm 17.9}^{\ddagger}$ | $43.0 \pm 8.0$ | $7.0 \pm 3.1$ | $1.0 \pm 0.0$ | $2.0 \pm 0.9$ |
| | Held-out, QSTConv | $\mathbf{12.3 \pm 5.2}$ | $3.7 \pm 1.2$ | $0.7 \pm 0.4$ | n/r | n/r |
| | Held-out, dropout | $\mathbf{6.6 \pm 4.6}$ | $2.3 \pm 2.0$ | $0.9 \pm 0.7$ | $0.6 \pm 0.2$ | $0.2 \pm 0.5^{\dagger}$ |
| E-VQC | Structure $\rho$ | 0.80 | 0.80 | 0.79 | 0.79 | 0.79 |
| | Train (%) | $25.8 \pm 2.3$ | $23.2 \pm 1.5$ | $12.8 \pm 9.1$ | $9.8 \pm 2.5$ | $8.8 \pm 2.0$ |

a third of unmatched pairs below 0.01 and lowers their mean to 0.05, with its learned logit scale at 24. Fixing the scale at 5 leaves 45% of them below 0.01 at epoch 200, and holding the kernel additive, which bounds the mean unmatched fidelity of a batch below by nearly $\frac{1}{2}$, still memorizes. The memorizing encoder therefore reaches the shattered configuration where the product kernel allows it, but does not need it to memorize.

**Backbones.** The two encoders order the heads identically and differ in how far they move each kernel. Without regularization QNervFormer fits every head more than QSTConv, and with branch dropout it still holds deep kernels away from the Haar value: at four layers its unmatched pairs average 0.029, 30 times $2^{-10}$, and 60% fall below 0.01, against $1.2 \times 10^{-3}$ and 99% on QSTConv. The encoder gradients follow. Under four-layer DR on QNervFormer the RMS encoder gradient grows from 0.021 at the first epoch to 0.17 at epoch 100 and 0.64 at epoch 200, and runs spend 3.4 of 9 checkpoints on the plateau. On QSTConv it decays from 0.042 to 0.022 and 0.019, and every checkpoint of every run sits on the plateau. The same flexibility lets unregularized QNervFormer shatter under QPMeL, and dropout on its convolution branches returns it to QSTConv's regime: its QPMeL kernel (0.50 matched, 0.25 unmatched, 0.2% below 0.01) then resembles QSTConv's (0.62, 0.34, 0.2%) rather than its own unregularized one.

Table F.4: *Trained kernels on THINGS-EEG2: mean fidelity of matched and of unmatched pairs, and the fraction of unmatched pairs below* 0.01*, on 1024 training pairs at the checkpoint of Table 2; five seeds (*$^{\dagger}$*four).* $^{\ddagger}$*QNervFormer without branch dropout at epoch 200, three seeds. Haar-random states give* $9.8 \times 10^{-4}$*,* $9.8 \times 10^{-4}$ *and* 1.00*. n/r, not run.*

| | | QSTConv | | | QNervFormer, branch dropout | | |
|---|---|---|---|---|---|---|---|
| Head | $L$ | Matched | Unmatched | $< 0.01$ | Matched | Unmatched | $< 0.01$ |
| E-VQC, fidelity | 4 | 0.057 | 0.034 | 0.39 | 0.127 | 0.072 | 0.18 |
| DR | 1 | 0.282 | 0.161 | 0.02 | 0.214 | 0.121 | 0.22 |
| DR | 2 | 0.188 | 0.117 | 0.06 | 0.123 | 0.072 | 0.28 |
| DR | 4 | $1.3 \times 10^{-3}$ | $1.2 \times 10^{-3}$ | 0.99 | 0.047 | 0.029 | 0.60 |
| DR | 8 | n/r | n/r | n/r | 0.019 | $8.9 \times 10^{-3}$ | 0.87 |
| DR | $16^{\dagger}$ | n/r | n/r | n/r | $9.7 \times 10^{-4}$ | $9.7 \times 10^{-4}$ | 1.00 |
| QPMeL | – | 0.625 | 0.336 | 0.002 | 0.501 | 0.252 | 0.002 |
| QPMeL, no dropout$^{\ddagger}$ | – | – | – | – | 0.557 | 0.053 | 0.32 |

### F.4 MS-COCO

MS-COCO Captions (Chen et al., 2015) runs QPMeL at 20 qubits with pretrained Xception (Chollet, 2017) and BERT (Devlin et al., 2019) encoders, trained with the symmetric InfoNCE objective on 30,000 training images and their captions. Table F.5 reports cross-modal matching on training pairs in both directions, for the closed-form PMeF and for statevector simulation of the 20-qubit circuit, with confidence intervals over episodes. QPMeL matches an image to its caption among 15 candidates with $94.7\%$ accuracy and a caption to its image with $95.9\%$ (chance $6.7\%$), and the two kernels agree within 3 points at every way count. The accuracies are measured on training pairs, so they show fit at 20 qubits, not generalization.

Table F.5: *MS-COCO cross-modal matching on training pairs (%), 20 qubits: closed-form PMeF and statevector simulation of the same circuit. Chance is* $100/n$ *for* $n$*-way.*

| | Image → text | | Text → image | |
|---|---|---|---|---|
| Way | PMeF | Simulated | PMeF | Simulated |
| 3 | $99.6 \pm 0.6$ | $98.7 \pm 1.9$ | $98.9 \pm 1.0$ | $98.0 \pm 2.3$ |
| 5 | $98.5 \pm 0.9$ | $98.8 \pm 1.4$ | $98.5 \pm 0.8$ | $98.0 \pm 1.7$ |
| 7 | $97.8 \pm 0.9$ | $96.3 \pm 2.1$ | $97.5 \pm 1.1$ | $96.3 \pm 2.4$ |
| 9 | $98.1 \pm 0.8$ | $95.3 \pm 2.1$ | $97.0 \pm 1.0$ | $95.1 \pm 2.4$ |
| 11 | $96.5 \pm 0.9$ | $94.4 \pm 2.1$ | $96.8 \pm 0.8$ | $95.1 \pm 1.7$ |
| 13 | $95.8 \pm 0.9$ | $95.2 \pm 1.8$ | $95.6 \pm 0.9$ | $93.7 \pm 2.2$ |
| 15 | $94.7 \pm 1.0$ | $93.6 \pm 1.9$ | $95.9 \pm 0.9$ | $93.1 \pm 1.7$ |

### F.5 Simulated and Executed Circuits: Inversion vs. SWAP Test

The noiseless and noisy simulations do not run the circuit that the physical QPU runs. Both estimate the same quantity, the product-state fidelity $|\langle\psi(x)|\psi(y)\rangle|^2$, but through two different primitives with very different hardware cost. We state the discrepancy explicitly because it determines how the hardware numbers in Table 3 should be read.

**Simulation: the inversion test.** The simulator prepares $|\psi(x)\rangle$ on $N$ qubits and then applies the inverse preparation for $|\psi(y)\rangle$, measuring the probability of returning to the all-zeros string, $P(|0\rangle^{\otimes N}) = |\langle\psi(y)|\psi(x)\rangle|^2$. Because both states are products, the circuit is $R_Y(2\theta_q^x)$, $R_Z(\gamma_q^x)$, $R_Z(-\gamma_q^y)$, $R_Y(-2\theta_q^y)$ on each qubit independently: width $N$, depth 4, and *zero* two-qubit gates. Transpiled against the IBM heavy-hex basis this remains depth 6 with no entangling gates at every qubit count tested.

**Hardware: the SWAP test.** The QPU runs a SWAP test instead. It uses $2N + 1$ qubits, one control plus two $N$-qubit registers holding $|\psi(x)\rangle$ and $|\psi(y)\rangle$, applies a Hadamard on the control, $N$ controlled-SWAP gates between corresponding register qubits, a second Hadamard, and measures only the control. The fidelity is recovered as $2P(0) - 1$. The historical implementation used this primitive with `SamplerV2` parameter binding and one-bit readout.

**Cost of that choice.** The SWAP test is not free. Each CSWAP decomposes into 8 CNOT-equivalent two-qubit gates, so the logical circuit carries $8N$ entangling gates where the inversion test carries none, and heavy-hex routing adds more. Table F.6 reports transpiled counts for both primitives against the `ibm_marrakesh` coupling map at optimization level 1.

At $N = 10$ the recorded transpilation uses roughly twice the qubits, 167 two-qubit gates instead of none, and about $65\times$ the depth of the inversion circuit. The comparison between the 3000-shot simulation and the 128-shot hardware run confounds shot noise, device noise, and the estimator circuit.

Table F.6: *Transpiled cost of the two fidelity primitives, targeting the* `ibm_marrakesh` *coupling map and basis at optimization level* 1. *Two-qubit counts are native* `cz` *gates after routing; the pre-routing logical count is* $8N$.

| **Primitive** | **Setting** | **Qubits** | **2-qubit gates** | **Depth** |
|---|---|---|---|---|
| Inversion test | $N = 5$ | 5 | 0 | 6 |
| Inversion test | $N = 6$ | 6 | 0 | 6 |
| Inversion test | $N = 10$ | 10 | 0 | 6 |
| SWAP test | $N = 5$ | 11 | 73 | 187 |
| SWAP test | $N = 6$ | 13 | 93 | 232 |
| SWAP test | $N = 10$ | 21 | 167 | 394 |