Investigating barren plateaus in classically trained IQP generative models

Hela Mhiri, Koshik Seeburrun, Mohamed Yaqub Alameen, and Alejandro Jackson

June 30, 2026

Report for Advanced Quantum Information Long Project
Master QI , Sorbonne Université

1 Introduction

1.1 Variational quantum algorithms

Variational quantum algorithms (VQAs) use a parameterized quantum circuit together with a classical optimizer. The circuit prepares a state that depends on parameters θ, a loss function is estimated from measurement outcomes or expectation values, and the optimizer updates θ. The hope is that the circuit gives a useful model class while the classical optimizer handles the search over parameters. [1]

1.1.1 General framework

A typical VQA has three parts:

$$\theta \longrightarrow U(\theta) \longrightarrow \mathcal{L}(\theta) \longrightarrow \theta_{\text{new}}.$$

Here $U(\theta)$ is the parameterized circuit and $\mathcal{L}(\theta)$ is the training objective. In this project the objective is an MMD loss between a data distribution p and an IQP model distribution q_θ. The main trainability question is whether the gradient $\nabla_\theta \mathcal{L}$ remains large enough to guide optimization as the number of qubits grows.

1.1.2 Quantum generative modeling

In quantum generative modeling, a parameterized circuit is trained so that its measurement outcomes look like samples from a target distribution. If the circuit prepares $|\psi(\theta)\rangle$, then the model probability of a bitstring x is

$$q_\theta(x) = |\langle x|\psi(\theta)\rangle|^2.$$

We train with Maximum Mean Discrepancy (MMD), an implicit loss that compares distributions through sample statistics rather than exact likelihoods. For IQP circuits, the relevant statistics can be written as parity observables, or Pauli Z-word expectations.

1.1.3 Barren Plateaus

A barren plateau is a region of the loss landscape where gradients concentrate near zero. In scaling language, the warning sign is

$$\mathrm{Var}_\theta\left[\partial_{\theta_j}\mathcal{L}(\theta)\right] \sim C\,2^{-\alpha n}$$

for some $\alpha > 0$. Then most random parameter settings give almost no useful update direction. This matters even when training is classical: exact gradients are not very helpful if the landscape itself has lost signal. We therefore study gradient concentration as a function of circuit family, initialization, and MMD bandwidth.

For deep randomly initialized circuits, the average gradient is typically zero due to symmetry in the parameter space. Consequently, trainability is determined by the gradient variance.

A large variance indicates that the loss landscape contains meaningful slopes that can guide optimization. In contrast, a barren plateau occurs when the variance decreases exponentially with the number of qubits.

1.2 IQP circuits

Instantaneous Quantum Polynomial-time (IQP) circuits are a class of commuting quantum circuits. A standard IQP circuit starts in $|0\rangle^{\otimes n}$, applies Hadamard gates to create a uniform superposition, applies a diagonal commuting layer, and then applies Hadamard gates again before measurement:

$$|\psi(\theta)\rangle = H^{\otimes n} D_\theta H^{\otimes n} |0\rangle^{\otimes n}.$$

The diagonal layer D_θ is parameterized by interaction terms. In this project, those interactions are represented by a binary generator matrix $G \in \{0,1\}^{m\times n}$. Each row g_j of G is a binary mask selecting the qubits touched by one generator, and the parameter θ_j controls the strength of that generator. Equivalently, the rows of G define a hypergraph: qubits are vertices, and generators are hyperedges.

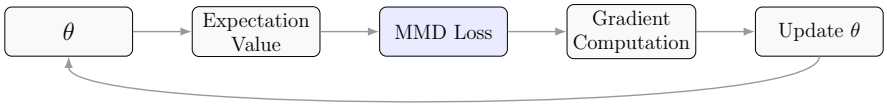
The generator matrix is the main structural object in the project. Its row weight gives the number of qubits in one interaction, its overlaps describe how generators share qubits, and the rule for constructing G as n changes defines an IQP circuit family. The primary families considered are the product-state family, the ZZ lattice family, the sparse Erdos-Renyi family, and the complete-graph family. These families let us compare no-entanglement, local, random sparse, and all-to-all pairwise connectivity regimes.

IQP circuits are important here for two reasons. First, sampling from general IQP output distributions is believed to be classically hard under standard complexity-theoretic assumptions, which makes them interesting candidates for quantum generative modeling. Second, the same commuting structure that makes

the circuits special also gives closed-form expressions for the expectation values used in MMD training. This creates the central tension of the project: training can be performed classically through parity and Z-word expectation formulas, while final sampling from the trained distribution can still be assigned to a quantum device.

1.2.1 IQP circuits for generative modeling

IQP Circuits do not necessarily need to run quantum circuit during training. We can train it classically using the following pipeline,



Parameters θ in an IQP circuit control certain quantum gates, hence the entire model is just a vector of numbers. The main optimization held here is by calculating and minimizing the MMD loss. MMD is computed using expectations of functions over the distribution and the loss can be termed as the average similarity between data samples.

Generally, sampling from a quantum circuit requires quantum runs. But for IQP circuits these expectation values can be computed analytically, because IQP circuits' probability distribution can be expressed using Fourier-type expressions. This lets them compute the necessary statistics using classical sums.

Why do we call it classical? Because the quantities inside MMD reduce to expressions like: Expectation of parity functions or Expectation of products of bits and for IQP circuits these expectations have closed-form formulas. Hence, instead of simulating a full quantum state for an exponential cost, we compute a structured formula for a polynomial cost.

1.3 Overview of results

The paper by Recio-Armengol, Ahmed, and Bowles gives the starting point for this project [2]. It shows that IQP generative models trained with MMD can be optimized on classical hardware: the MMD loss reduces to Pauli Z-word expectation values, and those expectations can be estimated efficiently for IQP circuits. They train models with up to one thousand qubits, use data-dependent initialization to avoid the barren-plateau problems seen in many variational circuits, and still reserve the final sampling step for quantum hardware.

Our experiments make that picture more conditional. We did not find a simple hypergraph-only explanation for barren plateaus. Instead, our empirical benchmarks reveal a strict rule involving the interplay of initialization, system size, and circuit structure. We demonstrate that uniform initialization was the most problematic, most prone to suffering from barren plateaus, while small-angle initialization performed slightly better before also failing as the system size scaled. Data-dependent variations worked best to prevent barren plateaus, with the layer-wise approach demonstrating viable scaling capabilities, at the cost of being highly time-consuming (layer-by-layer training). We also found that final MMD loss can hide failures in marginal agreement, especially when the bandwidth changes which correlation orders the loss sees. For this reason, we treat gradient behavior, target-relative concentration, and marginal error as separate diagnostics rather than reducing the result to one pass/fail number.

2 Training procedure

2.1 Datasets

The Datasets are particularly generated to follow a certain format of bitstrings (0/1 vectors) from different sources:

1. **MNIST** - Consists of handwritten images of digits in binary (0 or 1) by flattening each image into a long vector of pixels and converting each pixel's value to a range of [0,1].

$$f(\text{pixel}) = \begin{cases} 1 & \text{if pixel} > 0.5 \\ 0 & \text{if pixel} \leq 0.5 \end{cases}$$

2. **Ising** - Generates samples from a physical system, resembling a configuration of magnetic spins, converting values to {-1, +1} and further in binary. Each individual value of spins recorded is not random but follows a complex physical distribution.

3. **BLOBS** - Creates fake synthetic data by generating noisy variations and multiple clusters too. This brings a multi-modal distribution by providing noise addition to a central pattern in each sample.

4. **Genomic** - Real-life biological complex distribution possessing DNA variations encoded as binary values.

2.2 Training classically

2.2.1 MMD Loss

The model is trained by reducing the MMD loss between the data distribution and the distribution generated by the parameterized IQP circuit. The main function compares properties of the real data and the model distribution instead of reconstructing the full probability distribution entirely. The MMD loss is given by :

$$\mathcal{L}_{\text{MMD}} = \mathbb{E}_{x,x' \sim p_{\text{data}}}[k(x,x')] + \mathbb{E}_{y,y' \sim p_\theta}[k(y,y')] - 2\mathbb{E}_{x \sim p_{\text{data}},\, y \sim p_\theta}[k(x,y)],$$

where p_{data} is the data distribution, p_θ is the model distribution from IQP circuit, and $k(\cdot,\cdot)$ is kernel function measuring common features between samples.

Instead of evaluating the full quantum probability distribution, the training procedure only requires specific expectation values of functions such as parity functions. For a subset of qubits S, a parity feature is defined as

$$f_S(x) = (-1)^{\sum_{i \in S} x_i}.$$

Later, the loss and its gradients can be evaluated completely on a classical computer without simulating the full 2^n dimension quantum state.

2.2.2 Training and Test Loss Behaviour

The plot compares the training and test loss over optimization iterations for two independent runs. In both cases, the loss decreases rapidly in the early stages and then gradually saturates, indicating convergence of the MMD loss. After convergence, both losses stabilize close to zero, indicating that the model distribution closely matches the target data distribution.

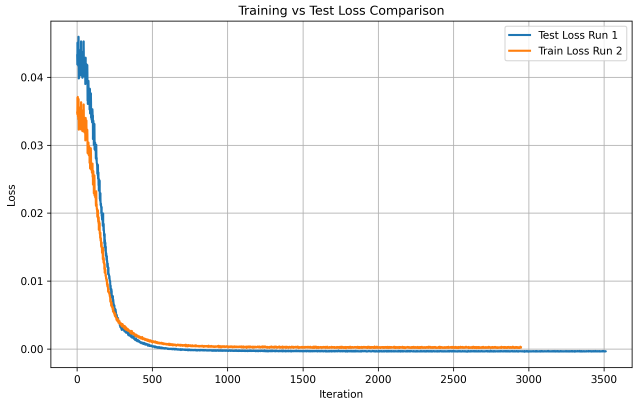


Figure 1: Training and test loss comparison over optimization iterations.

2.2.3 Connectivity and Initialization

The circuit structure is defined from `gates_config` hyperparameter. We use `name:"local_gates"` which creates local multi-qubit interaction terms in the IQP circuit.

For graph-structured datasets, the configuration name:"nearest_neighbor_gates" is used. In these datasets, gates are placed only between neighboring nodes of a predefined graph, allowing circuit connectivity to follow the topology of dataset.

The trainable parameters θ are initialized using the `init_config` settings. Standard initialization uses small random values:

$$\theta_i \sim \mathcal{N}(0, \sigma^2),$$

where $\sigma =$ `init_scale`. Small values such as `init_scale` $= 10^{-4}$ initialize the circuit close to the identity transformation and improve stable optimization during early stages of training the system. Additional randomness can also be added through `param_noise`, which concerns the initialized parameters with small Gaussian noise.

For example `init_scale`= -1, activates data-dependent initialization instead of standard random initialization.

3 Initialization strategies

In IQP circuits training is fundamentally bounded by the optimization landscape. As circuit depth and system size (n) increase, parameterized quantum circuits are prone to the Barren Plateau phenomenon. This exponentially flat landscape renders classical optimizers ineffective. The objective of an initialization strategy is to mathematically steer the initial state of the quantum circuit into a favorable region with sufficient gradient variance before the training process even begins.

3.1 Uniform initialization

Uniform initialization involves assigning initial circuit parameters θ from a broad, uniform distribution, typically across the full range of the gate operators. While this provides a stochastic starting point, it induces high levels of entanglement within the initial quantum state which leads to the collapse of gradient variance.

Previous theoretical works predicted an exponential gradient collapse in expressive variational circuits [10]. Empirically, Figure 2 confirms this behavior, showing a clear drop in gradient variance from nearly 10^{-1} down to 10^{-3} as the system scales from $n = 4$ to $n = 25$ qubits.

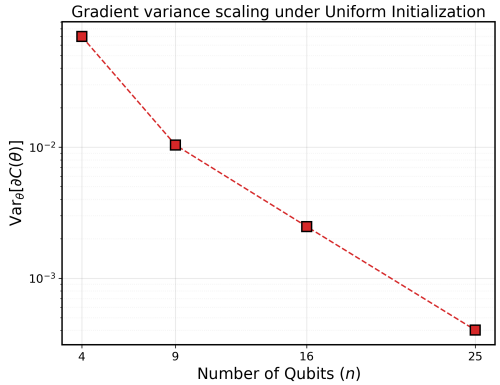


Figure 2: Gradient variance under uniform initialization. As the number of qubits increases, the variance decays exponentially, this demonstrates barren plateau behavior in IQP-MMD training.

3.2 Small angle initialization

Small angle, or identity, initialization mitigates initial entanglement by constraining parameters θ to values at or near zero, ensuring that the quantum gates act as identity operations thus preventing the circuit from immediately entering a highly concentrated region of Hilbert space [12]. This approach provides a baseline that maintains high gradient variance for shallow circuits. However, as the circuit depth increases to improve

expressivity, entanglement progressively accumulates throughout the circuit. Eventually, the gradients again begin to concentrate near zero, especially at larger qubit counts.

3.3 Data dependent initialization

Data-dependent initialization encodes the statistical features of the target dataset directly into the quantum state by mapping classical covariance patterns to the parameters of two-qubit interacting gates. By setting these parameters based on the training data, the circuit receives a "warm start" that aligns the output distribution with the independent average of the input, effectively positioning the model near a favorable local minimum [11]. This methodology preserves high gradient variance for shallow circuits.

A patch is a localized neighborhood of data points. The radius r determines the size of this neighborhood. A small r encodes only on local relationships, while a large r attempts to capture long-range, complex patterns into the circuit.

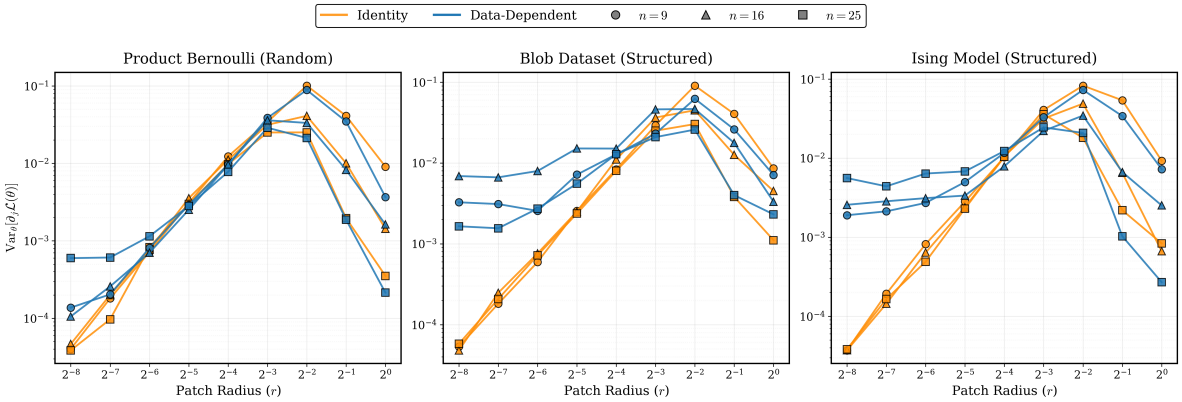


Figure 3: Compares identity initialization and data-dependent initialization across increasing system sizes.

Figure 3 demonstrates that on structured data (Ising and Blobs), data-dependent initialization consistently preserves the gradient variance better than identity initialization at larger system sizes ($n = 25$). While this strategy successfully delays barren plateau formation, the variance still exhibits a gradual degradation for sufficiently large systems. For the random dataset, the identity and data dependent initialization are somewhat equivalent because random noise contains no real statistical correlations, the data-dependent algorithm extracts no meaningful pattern.

3.4 Layer-wise initialization

Layer-wise initialization addresses the scaling limitations in the data-dependent strategy by employing an iterative expansion of the quantum circuit. This method starts with a single-layer circuit optimized via data-dependent initialization. Additional layers are then integrated with their parameters initialized to zero [12]. Because these new layers initially function as identity operators, the expanded circuit corresponds to the preceding training step while providing the necessary capacity for further expressivity. This incremental approach allows the model to progressively capture more complex data distributions while maintaining a stable gradient.

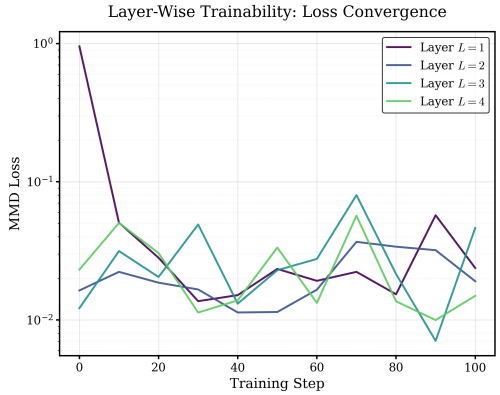


Figure 4: Layer-wise trainability on the Ising dataset using a 2D lattice IQP architecture.

Figure 4 shows the trajectory of layer-wise training ($L = 1$ through $L = 4$), progressively adding and training layers sequentially keeps the training MMD loss stable within a highly trainable window between 10^{-2} and 10^{-3} without flattening into a barren plateau. This confirms that layer-wise initialization is a highly effective strategy to bypass barren plateaus and can scale the model to handle deeper complexity.

A complete graph structure was also tested. However, the all-to-all connectivity of a complete graph immediately caused heavy entanglement. This made the training process exceptionally time-consuming due to current hardware limitations, rendering it impractical as a scaling solution.

4 Inference procedure

Our inference procedure examines two diagnostics that provide insight as to the potential presence of barren plateaus in IQP circuits. The first is *marginal agreement*: whether the learned IQP distribution q_θ matches the target distribution p on subsets of variables. The second is *anti-concentration*: how spread out a single learned distribution is over $\{0,1\}^n$. These answer different questions. Marginals test whether the model learned the target; anti-concentration tests the shape of the learned distribution.

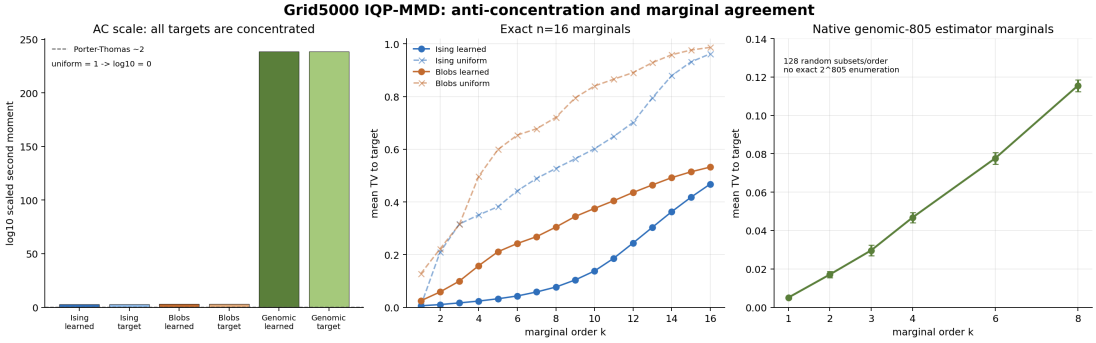


Figure 5: Anti-concentration and marginal diagnostics. Left: target-relative concentration. Middle: exact $n = 16$ marginal TV for Ising and blobs. Right: sampled low-order marginal TV for genomic-805.

For small systems, we compute the learned distribution exactly. Given an IQP circuit, we enumerate all $z \in \{0,1\}^n$, form the diagonal phase vector

$$d(z) = \exp(-i\phi(z)), \qquad \phi(z) = \sum_j \theta_j (-1)^{z \cdot g_j},$$

apply a fast Walsh–Hadamard transform, and square the resulting amplitudes:

$$q_\theta(x) = \left| 2^{-n} \sum_{z \in \{0,1\}^n} (-1)^{x \cdot z} d(z) \right|^2 .$$

This exact route is exponential in n, so it is used only for small systems. For native large-n datasets, such as the genomic-805 experiment, exact enumeration is impossible; there we use estimator-based diagnostics on sampled Pauli observables and sampled low-order subsets.

The main reason to stratify diagnostics by marginal order is that the Gaussian-kernel MMD loss decomposes in the Walsh basis as

$$\mathrm{MMD}^2_\sigma(p, q_\theta) = C_\sigma \sum_{S \subseteq [n]} \tau^{|S|} \left(\langle Z_S \rangle_p - \langle Z_S \rangle_{q_\theta}\right)^2, \qquad \tau = \tanh\left(\frac{1}{4\sigma^2}\right).$$

For small n we compute q_θ exactly. For larger n, exact enumeration is infeasible, so anti-concentration is treated as an estimator diagnostic. In the Pauli-estimator route, the scaled second moment is estimated through a Parseval identity using random non-identity Pauli strings:

$$\widehat{S} = 1 + (2^n - 1)\,\overline{Y}, \qquad Y = \langle Z_a \rangle^2_{q_\theta}.$$

The reason to report marginals by order is the Gaussian MMD decomposition

$$\mathrm{MMD}^2_\sigma(p, q_\theta) = C_\sigma \sum_{S \subseteq [n]} \tau^{|S|} \left(\langle Z_S \rangle_p - \langle Z_S \rangle_{q_\theta}\right)^2, \qquad \tau = \tanh\left(\frac{1}{4\sigma^2}\right).$$

The bandwidth σ decides which correlation orders the loss pays attention to. Large bandwidths weight low-order structure more heavily; smaller bandwidths put more mass on higher-order terms and can make training noisier.

4.1 Assessing marginals

For each subset S of size k, we compare the target marginal p_S with the learned marginal $(q_\theta)_S$ using

$$\mathrm{TV}(p_S, (q_\theta)_S) = \frac{1}{2} \sum_{u \in \{0,1\}^{|S|}} \left|p_S(u) - (q_\theta)_S(u)\right|.$$

We then average over subsets of the same order:

$$T_k = \mathbb{E}_{|S|=k}[\mathrm{TV}(p_S, (q_\theta)_S)].$$

We also track the averaged squared Z-word error F_k, since that is the term appearing inside the MMD objective. T_k measures distributional mismatch on k-variable marginals, while F_k measures the squared parity-correlation error seen by the Gaussian-kernel MMD objective. This matters because a small kernel-weighted MMD can hide marginal errors at orders that the chosen bandwidth weights weakly. A low final MMD value can still hide bad marginal agreement. In the bandwidth sweep, the small-bandwidth genomic-805 run reached a low loss but had much worse low-order marginal TV than the paper and large-bandwidth runs.

On the exact $n = 16$ controls, the paper bandwidth gave the best high-order marginal behavior for Ising and blobs, while the large bandwidth mainly helped on easier low-order checks. The larger error seen at small bandwidth is not enough on its own to identify the cause of failure. It could be a classical ML issue, such as overfitting to a narrow kernel objective, or a quantum trainability issue, such as barren-plateau-like gradient collapse; separating these explanations requires a closer analysis of gradients, initialization, and marginal behavior.

4.2 Assessing anti-concentration

For anti-concentration we use the scaled second moment

$$S(q_\theta) = 2^n \sum_x q_\theta(x)^2,$$

with effective support $2^n/S(q_\theta)$. Uniform has $S = 1$; more concentrated distributions have larger S. For large n, S is estimated with sampled Pauli strings through the Parseval estimator

$$\widehat{S} = 1 + (2^n - 1)\,\overline{Y}, \qquad Y = \langle Z_a \rangle^2_{q_\theta}.$$

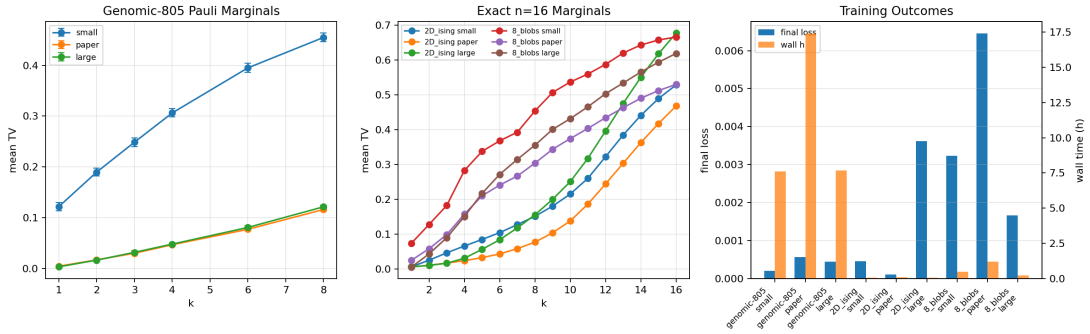


Figure 6: Bandwidth sweep for marginal agreement. Small bandwidth gives poor genomic-805 low-order TV despite low loss; on exact $n = 16$ controls, the paper bandwidth is strongest for high-order Ising and blobs marginals.

These estimator rows are not exact certificates; they come with Monte Carlo budget and uncertainty.

We do not reduce anti-concentration to a pass/fail flag. The threshold $S \geq 1$ is vacuous, and a learned distribution should be compared with the target, not with uniform alone. In our runs, learned IQP-MMD distributions are not Porter–Thomas-like or uniform-scale anti-concentrated. That is not automatically bad: the targets are concentrated too. The exact Ising learned and target second moments are on the same scale, blobs is more concentrated than the learned model, and genomic-805 is extreme because a few thousand observed haplotypes sit inside a 2^{805} space.

The final diagnostic is therefore the joint picture

$$\{S(q_\theta),\ S(p) - S(q_\theta),\ T_k,\ F_k\}_k\,.$$

This keeps two issues separate: whether the learned distribution has the right concentration scale, and whether it matches the target marginals that the MMD loss is supposed to learn.

5 Conclusion

IQP-MMD training is promising because the expensive part of variational quantum learning can be moved to classical computation. Our results suggest that this does not remove trainability questions entirely. Barren-plateau behavior depends on initialization, circuit structure, and system size, and a low final MMD loss does not always mean the learned distribution matches the target marginals. The safest conclusion is narrow: IQP circuits can be trained classically in useful regimes, but their success should be checked with gradient scaling, marginal agreement, and target-relative anti-concentration rather than with one loss value.

References

[1] M. Cerezo, A. Arrasmith, R. Babbush, S. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles. *Variational quantum algorithms*. Nature Reviews Physics, 3(9):625–644, 2021. arXiv:2012.09265.

[2] E. Recio-Armengol, S. Ahmed, and J. Bowles. *Train on classical, deploy on quantum: scaling generative quantum machine learning to a thousand qubits*. arXiv:2503.02934, 2025.

[3] XanaduAI. *scaling-gqml*. GitHub repository, `github.com/XanaduAI/scaling-gqml`, accessed 2026-05-26.

[4] alejack312. *iqp-mmd-barren-plateau*. GitHub repository, `github.com/alejack312/iqp-mmd-barren-plateau`, accessed 2026-05-26.

[5] M. S. Rudolph et al. *Trainability barriers and opportunities in quantum generative modeling*. arXiv:2305.02881, 2023.

[6] M. Larocca et al. *Barren plateaus in variational quantum computing*. Nature Reviews Physics, 2024.

[7] M. J. Bremner, A. Montanaro, and D. J. Shepherd. *Average-case complexity versus approximate simulation of commuting quantum computations*. arXiv:1504.07999, 2015.

[8] M. J. Bremner, A. Montanaro, and D. J. Shepherd. *Achieving quantum supremacy with sparse and noisy commuting quantum computations*. arXiv:1610.01808, 2016.

[9] S. Herbst, I. Brandic, and A. Perez-Salinas. *Limits of quantum generative models with classical sampling hardness*. arXiv:2512.24801, 2025.

[10] M. Larocca, P. Czarnik, A. D. Alking, and M. Cerezo. *Diagnosing barren plateaus with tools from quantum optimal control*. Quantum, 6:824, 2022.

[11] E. Fontana et al. *A unifying account of warm start guarantees for patches of quantum landscapes*. arXiv:2502.07889v1 [quant-ph], 2025.

[12] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles. *A Review of Barren Plateaus in Variational Quantum Computing*. arXiv:2405.00781 [quant-ph] (Updated Review), 2025.