

Trainability and Mode Separation of Mixed IQP-QCBMs

Youngseok Lee and Hyunwoo Kim

Quantum AI Team, NORMA Inc., Republic of Korea

ys_lee@norma.co.kr, hw_kim@norma.co.kr

1 Introduction

The IQP-QCBM, a quantum circuit Born machine (QCBM) [1, 2] built from an instantaneous quantum polynomial-time (IQP) circuit of commuting Pauli- X rotations, is a promising quantum generative model for the NISQ era: it can be trained classically, while sampling from the trained model is believed to be classically intractable unless the polynomial hierarchy collapses [4, 5]. Its maximum mean discrepancy (MMD) loss and the loss gradients are assembled from low-body Pauli- Z expectation values that admit efficient classical estimates [3]; the entire training therefore runs classically, and quantum hardware is needed only to sample from the trained model.

However, the model faces a tension between *trainability* and *expressivity*. The ancilla-free IQP-QCBM admits initializations that provably avoid barren plateaus [6–8], but it is not universal [9]: there exist target distributions it cannot represent. The ancilla-augmented IQP-QCBM, in contrast, becomes universal with sufficiently many ancilla qubits, but whether its training can still avoid barren plateaus remains unknown.

Here, we address this tension by proposing the *mixed IQP-QCBM*, a generalization of the ancilla-augmented IQP-QCBM that keeps its expressivity and is provably trainable. The mixed IQP-QCBM decomposes an ancilla-augmented IQP circuit into $L = 2^a$ independent ancilla-free IQP circuits, where a is the number of ancilla qubits (Fig. 1). For this model we establish three initialization strategies: (i) data-agnostic, (ii) global data-dependent, and (iii) clustered data-dependent, which we find to be trainable. As experimental corroboration, we train the model on targets ranging from $n = 16$ Gaussian blobs to $n = 784$ MNIST images and find that adding ancillas consistently lowers the test MMD² loss across all benchmarks, with the successful models also exhibiting the mode separation predicted by our analysis.

2 The mixed IQP-QCBM: trainability and mode separation

We define the mixed IQP-QCBM as the weighted mixture

$$p_{\Theta}(z) = \sum_{\ell=0}^{L-1} \pi_{\ell} |\langle z | U(\theta^{(\ell)}) | \psi_0 \rangle|^2, \quad \sum_{\ell=0}^{L-1} \pi_{\ell} = 1, \quad (1)$$

of L independent ancilla-free IQP circuits $U(\theta^{(\ell)})$, called its *branches*, each carried with a weight π_{ℓ} (Fig. 1). When the weights are uniform, $\pi_{\ell} = 1/L$, a Walsh–Hadamard transform of the branch angles makes the mixture exactly equal to a single IQP circuit augmented with $\log_2 L$ ancilla qubits [10]; treating the weights π_{ℓ} as free trainable parameters therefore generalizes the ancilla-augmented IQP-QCBM.

We establish three trainable initializations of this model (Fig. 1(b)). For each, we prove that the mixture avoids barren plateaus by extending prior trainability results on ancilla-free IQP circuits [7, 8] to the ancilla-augmented setting. Each scheme sets only the single-qubit gate angles of the individual branches; the two-qubit gate angles start at zero.

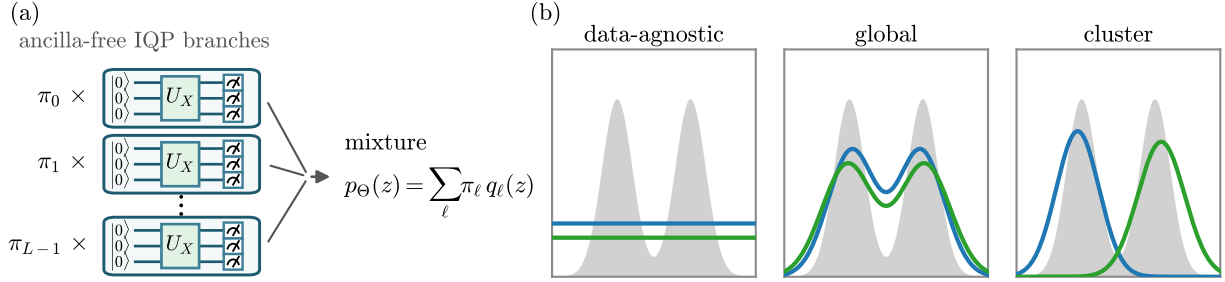


Figure 1: **(a)** The mixed IQP-QCBM: a mixture of $L = 2^a$ ancilla-free IQP branches, branch ℓ carried with weight π_ℓ ; with uniform weights the mixture equals a single IQP circuit augmented with a ancilla qubits. **(b)** The three trainable initializations for $L = 2$ over a faint bimodal target: *data-agnostic* (flat branches), *global* (branches nearly coincide on the full target), and *cluster* (each branch leans toward a distinct mode).

(i) *Data-agnostic*. Every branch starts at the unbiased center: all single-qubit gate angles are set to $\pi/4$, so the branches sample uniformly over the sample space, independently of the data; it is the barren-plateau-avoiding baseline of prior single-circuit work [7, 8].

(ii) *Global*. Every branch receives the same single-qubit gate angles on the visible (non-ancilla) qubits, fixed to match the data’s single-bit means, so the branches start coincident to each other.

(iii) *Cluster*. The data are first split into L clusters by spectral clustering under an RBF (Hamming–Gaussian) kernel; each branch is then initialized as in the global scheme but on its own cluster, so the branches start separated onto distinct parts of the data distribution.

We now state our main results:

Proposition 1 (Trainability, informal) *Under each of the three initializations above, the curvature of the MMD loss on a local patch around the starting point is supplied by the model’s own parameter sensitivity and does not vanish exponentially.*

Proposition 2 (Mode separation, informal) *The added ancillas improve the model only when the branches generate different distributions. To be precise,*

$$\mathcal{L}(\Theta) = \sum_{\ell} \pi_{\ell} \mathcal{L}_{\ell} - \overline{D}_{\text{branch}}^2, \quad \overline{D}_{\text{branch}}^2 = \frac{1}{2} \sum_{\ell, \ell'} \pi_{\ell} \pi_{\ell'} \sum_A w_A (z_A^{(\ell)} - z_A^{(\ell')})^2, \quad (2)$$

where the inter-branch diversity $\overline{D}_{\text{branch}}^2$ measures how differently the branches are distributed ($z_A^{(\ell)}$ is the expectation value of the Pauli-Z string $\prod_{i \in A} Z_i$ on a small set A of qubits in branch ℓ , and w_A its kernel weight). Moreover, the gradient of the ancilla couplings—the gates that build this diversity—is set by the initial value of $\overline{D}_{\text{branch}}^2$.

Proposition 2 sets a design goal for the initialization: supply branch diversity from the start. The data-agnostic and global schemes start with branches that differ only by a small initialization noise, so the initial diversity is near zero and must build up slowly during training. On the other hand, the cluster scheme starts with a diversity far above this noise scale, giving the ancilla couplings their gradient immediately.

3 Experiments and conclusion

Our experiments support both propositions from $n = 16$ to $n = 784$ (Fig. 2): consistent with Proposition 1, all three initializations train from the start on every benchmark. In particular, the

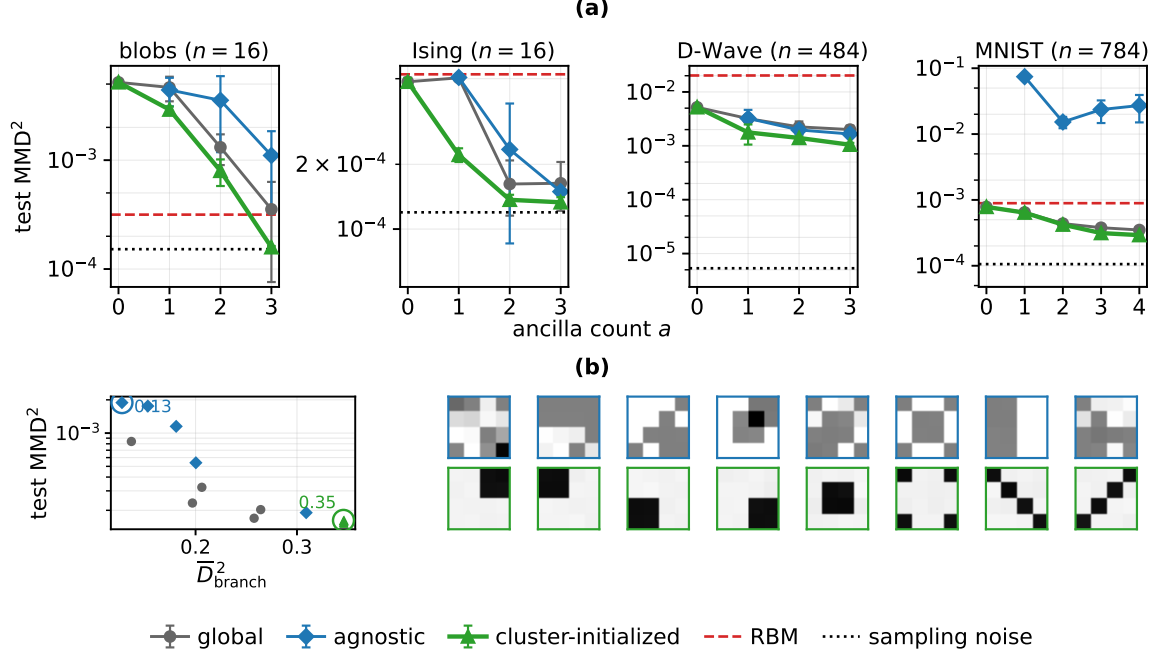


Figure 2: **(a)** Test MMD^2 (vertical, log scale) vs. ancilla count a ($L = 2^a$); mean $\pm$ one standard deviation over retrained seeds (five for $n = 16$, three otherwise). **(b)** Inter-branch diversity of the trained blobs models at $a = 3$. Left: $\overline{D}_{\text{branch}}^2$ (Eq. (2)), computed exactly from the per-branch Born distributions, against the test MMD^2 , one point per initialization and training seed. Right: the per-bit mean image of every branch of the two circled models (agnostic, top, $\overline{D}_{\text{branch}}^2 = 0.13$; cluster-initialized, bottom, 0.35).

cluster initialization we introduce lowers the test MMD^2 as ancillas are added throughout: on the small targets it reaches the level of the target’s own sampling noise, and on the two large targets, the $n = 484$ D-Wave data and $n = 784$ MNIST, it attains a test MMD^2 below that of a classical RBM.

On blobs at $a = 3$ the experiments bear out Proposition 2: with ancillas, successful training is accompanied by a growing inter-branch diversity. We compute $\overline{D}_{\text{branch}}^2$ exactly from the per-branch Born distributions of every trained model, five initialization seeds per scheme (Fig. 2(b)): the test MMD^2 falls as the diversity grows, and the cluster-initialized models reach the largest diversity and lowest loss—so reproducibly that their five points nearly coincide. The per-branch samples make this visible: the low-diversity branches produce near-interchangeable samples that mix several blob patterns, whereas every cluster-initialized branch generates a single distinct pattern. This is precisely the mechanism identified by Proposition 2: the added ancillas improve the fit by splitting the target across the branches, and an initialization that supplies this split from the start turns the added expressivity into a measurable gain.

We conclude this extended abstract with a note on future directions. Although the cluster scheme performs best across our benchmarks, its RBF clustering (Sec. 2) is only one way to route the branches, and a clustering better matched to the model may exist. The branches can also be strengthened: adding three- and four-qubit gates within a branch would raise its expressivity. Both are natural next steps.

References

- [1] M. Benedetti *et al.*, *Parameterized quantum circuits as machine learning models*, Quantum Sci. Technol. **4**, 043001 (2019).
- [2] J.-G. Liu and L. Wang, *Differentiable learning of quantum circuit Born machines*, Phys. Rev. A **98**, 062324 (2018).
- [3] M. S. Rudolph *et al.*, *Trainability barriers and opportunities in quantum generative modeling* (2024).
- [4] M. J. Bremner, R. Jozsa, and D. J. Shepherd, *Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy*, Proc. R. Soc. A **467**, 459 (2011).
- [5] M. J. Bremner, A. Montanaro, and D. J. Shepherd, *Average-case complexity versus approximate simulation of commuting quantum computations*, Phys. Rev. Lett. **117**, 080501 (2016).
- [6] J. R. McClean *et al.*, *Barren plateaus in quantum neural network training landscapes*, Nat. Commun. **9**, 4812 (2018).
- [7] K. Shen, S. Pielawa, V. Dunjko, and H. Wang, *Characterizing trainability of instantaneous quantum polynomial circuit Born machines*, arXiv:2602.11042 (2026).
- [8] S. Lerch, J. Bowles, R. Puig, E. Armengol, Z. Holmes, and S. Thanasilp, *IQP Born machines under data-dependent and agnostic initialization strategies*, arXiv:2603.14576 (2026).
- [9] A. Kurkin, K. Shen, S. Pielawa, H. Wang, and V. Dunjko, *Universality and kernel-adaptive training for classically trained, quantum-deployed generative models*, arXiv:2510.08476 (2025).
- [10] J. Slim, S. Monaco, F. Rehm, D. Krücker, and K. Borrás, *An IQP Born machine for calorimeter image generation at 64 qubits with compiled-IQP deployment*, arXiv:2605.27735 (2026).