

Supplementary Information for “Trainability barriers and opportunities in quantum generative modeling”

I. SUPPLEMENTARY NOTE - TECHNICAL NUANCES IN EXPLICIT AND IMPLICIT LOSSES

In this work, we have argued that there are two main classes of loss functions for generative modelling tasks. First are *explicit* losses which have the form

$$\mathcal{L}_{\text{expl}}(\boldsymbol{\theta}) := \sum_{\mathbf{x}_1 \dots \mathbf{x}_r} f\left(p(\mathbf{x}_1), \dots, p(\mathbf{x}_r), q_{\boldsymbol{\theta}}(\mathbf{x}_1), \dots, q_{\boldsymbol{\theta}}(\mathbf{x}_r)\right), \quad (1)$$

where $f(\cdot)$ is a function that depends on the target probabilities $p(\mathbf{x}_i)$ and model probabilities $q_{\boldsymbol{\theta}}(\mathbf{x}_i)$ for data variables $\mathbf{x}_i \in \mathcal{X}$ with $i = 1, \dots, r$ (but not the data samples themselves). The other are *implicit* loss functions which are expressed as

$$\mathcal{L}_{\text{impl}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_r \sim \{p, q_{\boldsymbol{\theta}}\}} g(\mathbf{x}_1, \dots, \mathbf{x}_r), \quad (2)$$

where $g(\mathbf{x}_1, \dots, \mathbf{x}_r)$ is some function that depends on the data (but not probabilities), and an expectation is over data variables $\mathbf{x}_1, \dots, \mathbf{x}_r$ sampled either from the data distribution p or the model distribution $q_{\boldsymbol{\theta}}$.

While almost all practical loss functions for generative models fall either into the definitions of explicit or implicit losses (see Sec. “Loss functions” in the main text), there exists a technical caveat which allows loss functions to be classified both as explicit and implicit. To provide a rather general example, we consider a general loss of the form

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{y}) g(\cdot) \quad (3)$$

$$= \mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\theta}}, \mathbf{y} \sim p} [g(\cdot)] \quad (4)$$

where the function g is for now left arbitrary. By comparing Eq. (3) and Eq. (1), we can see that for $\mathcal{L}(\boldsymbol{\theta})$ to be an explicit loss g cannot depend on $\mathbf{x}, \mathbf{y}$. Conversely, by comparing Eq. (4) and Eq. (2), for $\mathcal{L}(\boldsymbol{\theta})$ to be implicit, g cannot depend on $p(\mathbf{x})$ or $q_{\boldsymbol{\theta}}(\mathbf{y})$. Thus, it appears impossible for such a function to be both explicit and implicit. However, taking g as the Kronecker delta $\delta_{\mathbf{x}, \mathbf{y}}$ allows one to straddle the definitions. That is, for $g = \delta_{\mathbf{x}, \mathbf{y}}$, $\mathcal{L}(\boldsymbol{\theta})$ is implicit as it can still be written as an average over samples in Eq. (4) but also we can write $\mathcal{L}(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{x})$ which is of the form of an explicit loss given in Eq. (1).

This case is not just purely hypothetical and arises, for example, for the MMD loss, Eq. (13) in the main text, with the kernel $K(\mathbf{x}, \mathbf{y}) = \delta_{\mathbf{x}, \mathbf{y}}$. Here we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) \delta_{\mathbf{x}, \mathbf{y}} - 2 \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{y}) \delta_{\mathbf{x}, \mathbf{y}} + \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} p(\mathbf{x}) p(\mathbf{y}) \delta_{\mathbf{x}, \mathbf{y}} \\ &= \sum_{\mathbf{x} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})^2 - 2 q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{x}) + p(\mathbf{x})^2 \\ &= \sum_{\mathbf{x} \in \mathcal{X}} (q_{\boldsymbol{\theta}}(\mathbf{x}) - p(\mathbf{x}))^2, \end{aligned} \quad (5)$$

which takes the form of a pairwise explicit loss (see Eq. (3) in the main text). Consequently, the MMD is always an implicit loss function, but with this particular choice of kernel, it can additionally gain an explicit character. Interestingly, this loss is now very related to the TVD loss in Eq. (7) in the main text for which there exists no implicit form.

This example highlights that the distinction between explicit and implicit losses become non-mutually exclusive for implicit losses formulated using delta functions. Instead of adding this exception into either of the definitions of explicit or implicit losses, we acknowledge and embrace its existence. Through the steps outlined in Eq. (5), it may be possible to transform an explicit function exactly or approximately into a form which falls under the definition of an implicit function, which could then seemingly be used with implicit models. This may include functions whose Taylor series expansion in the probabilities converges quickly and can thus be approximated with a finite number of terms. Any term

with a positive integer power in the probabilities can then be transformed into an expectation over samples by leveraging the procedure outlined above. Unfortunately, we are not yet aware of any practical examples. In fact, we show that a function which can be classified as a pairwise explicit function directly suffers from untrainability with the conventional measurement strategy. There may however be cases where a quantum strategy can be found that leverages this δ -trick.

II. SUPPLEMENTARY NOTE - ANALYSIS ON PAIRWISE EXPLICIT LOSS FUNCTIONS

Here we detail the proofs of our no-go results regarding the pairwise explicit loss function. In particular, Theorem 1, which concerns the concentration of a statistical estimate of a pairwise explicit loss, is proved in Supplementary Note II A. In Supplementary Note II B, we show that a generative model with polynomial support but which has no inductive bias that aligns with the target distribution is untrainable. Finally, we use the Theorem 1 to derive Corollary 3 in Supplementary Note II C, showing the untrainability of the pairwise explicit loss.

For convenience, we recall relevant terms and notations. We are interested in training the model probabilities $q_\theta(\mathbf{x})$ to match some unknown target distribution $p(\mathbf{x})$ by minimizing the *pairwise explicit loss* of the form

$$\mathcal{L}(\theta) = \sum_{\mathbf{x}} f(p(\mathbf{x}), q_\theta(\mathbf{x})) , \quad (6)$$

where $f(\cdot)$ is some arbitrary function that measure the similarity between $p(\mathbf{x})$ and $q_\theta(\mathbf{x})$. As a reminder, this pairwise form encompasses the broad range of practical loss functions including the famous KL divergence, JS divergence, total variational distance as well as classical fidelity. See Sec. “Loss functions” in the main text for details.

Given the training dataset $\tilde{P}$ and model samples $\tilde{Q}_\theta$ corresponding to the empirical probabilities $\tilde{p}(\mathbf{x})$ and $\tilde{q}_\theta(\mathbf{x})$ respectively, the statistical estimate of the loss function can be written as

$$\tilde{\mathcal{L}}(\theta) = \sum_{\mathbf{x}} f(\tilde{p}(\mathbf{x}), \tilde{q}_\theta(\mathbf{x})) . \quad (7)$$

Crucially, we emphasise that, for the large system’s size, the number of training data and model samples can scale at most polynomially with the number of qubits i.e., $M = |\tilde{P}|, N = |\tilde{Q}_\theta| \in \mathcal{O}(\text{poly}(n))$.

A. Concentration of the pairwise explicit loss function: Proof of Theorem 1

In this section, we rigorously prove Theorem 1, which shows the concentration of the statistical estimate of the pairwise explicit loss. For convenience, the theorem is recalled below.

Theorem 1 (Concentration of pairwise explicit loss for concentrated models). *Consider the loss function of the form in Eq. (3) in the main text. Assume that for all bitstrings in the training dataset, $\mathbf{x} \in \tilde{P}$, the quantum generative model $q_\theta(\mathbf{x})$ exponentially concentrates towards some exponentially small value (as defined in Definition 1). Suppose that $N \in \mathcal{O}(\text{poly}(n))$ samples are collected from the quantum model corresponding to the set of sampled bitstrings $\tilde{Q}_\theta$, and that the training dataset $\tilde{P}$ contains $M \in \mathcal{O}(\text{poly}(n))$ samples. We define the fixed point of the loss as*

$$\mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta) = \sum_{\mathbf{x} \in \tilde{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \tilde{Q}_\theta} f(0, \tilde{q}_\theta(\mathbf{x})) , \quad (8)$$

with $\mathcal{P}$ (and $\mathcal{Q}_\theta$) being a set of unique bitstrings in $\tilde{P}$ (and $\tilde{Q}_\theta$). Then, the probability that the estimated value $\tilde{\mathcal{L}}(\theta)$ is equal to $\mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)$ is exponentially close to 1, i.e.,

$$\Pr_{\tilde{Q}_\theta, \theta}[\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)] \geq 1 - \delta , \quad (9)$$

with $\delta \in \mathcal{O}\left(\frac{\text{poly}(n)}{c^n}\right)$ for some $c > 1$.

Proof. The statistical estimate $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ is equal to $\mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta) = \sum_{\mathbf{x} \in \mathcal{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \mathcal{Q}_\theta} f(0, \tilde{q}_\theta(\mathbf{x}))$ when there is no overlap between $\tilde{P}$ and $\tilde{Q}_\theta$ i.e., $\tilde{P} \cap \tilde{Q}_\theta = \{\}$. The proof of the theorem is equivalent to proving that, after taking N measurement shots, the probability of not obtaining any bitstrings in the training dataset is exponentially close to 1. The probability that there is no overlap between $\tilde{P}$ and $\tilde{Q}_\theta$ is the probability that $\tilde{q}_\theta(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{P}$ which is also equivalent to $\Pr[\sum_{\mathbf{x} \in \mathcal{P}} \tilde{q}_\theta(\mathbf{x}) = 0]$ as $\tilde{q}_\theta(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{X}$. So, we have

$$\Pr_{\tilde{Q}_\theta, \theta}[\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)] = \Pr_{\tilde{Q}_\theta, \theta} \left[\sum_{\mathbf{x} \in \mathcal{P}} \tilde{q}_\theta(\mathbf{x}) = 0 \right] \quad (10)$$

$$= \int_0^1 \Pr_{\tilde{Q}_\theta} \left[\sum_{\mathbf{x} \in \mathcal{P}} \tilde{q}_\theta(\mathbf{x}) = 0 \middle| \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] \Pr_\theta \left[\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (11)$$

$$= \int_0^1 (1-s)^N \Pr_\theta \left[\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (12)$$

$$\geq \int_{\mu_s - \sqrt{\sigma_s}}^{\mu_s + \sqrt{\sigma_s}} (1-s)^N \Pr_\theta \left[\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (13)$$

$$\geq (1 - (\mu_s + \sqrt{\sigma_s}))^N \int_{\mu_s - \sqrt{\sigma_s}}^{\mu_s + \sqrt{\sigma_s}} \Pr_\theta \left[\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (14)$$

$$\geq (1 - (\mu_s + \sqrt{\sigma_s}))^N (1 - \sigma_s). \quad (15)$$

In the second equality Bayes' theorem is used to introduce the conditional probability of non-overlap between model samples and the training dataset for given $s = \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x})$ and the marginal probability is obtained by summing over all possible values of $\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x})$. The third equality uses the independence of each model sample and the fact that the probability that one drawn model bitstring is not in the training dataset is given by $(1-s)$. The first inequality is due to restricting the integration range (as the integrand is always greater than zero) with μ_s and σ_s^2 being the mean and variance of $\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x})$ over $\boldsymbol{\theta}$. The second inequality is due to taking the maximum value of s and thus the minimum value of $(1-s)^N$ within the integration range. To see how to reach the last line, we invoke Chebyshev's inequality on $\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x})$

$$\Pr_\theta \left[\left| \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) - \mu_s \right| \geq k\sigma_s \right] \leq \frac{1}{k^2}. \quad (16)$$

By specifying $k = 1/\sqrt{\sigma_s}$ and inverting the inequality, we have

$$\Pr_\theta \left[\left| \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) - \mu_s \right| \leq \sqrt{\sigma_s} \right] \geq 1 - \sigma_s. \quad (17)$$

With $\int_a^b \Pr[x]dx = \Pr[a \leq x \leq b]$, we then get Eq. (15).

To further bound the probability, we show that μ_s and σ_s^2 are exponentially small in the number of qubits. First, consider the mean

$$\mu_s = \mathbb{E}_\theta \left[\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) \right] \quad (18)$$

$$= \sum_{\mathbf{x} \in \mathcal{P}} \mu(\mathbf{x}) \quad (19)$$

$$\leq N_p \max_{\mathbf{x} \in \mathcal{P}} [\mu(\mathbf{x})], \quad (20)$$

with $N_p = |\mathcal{P}| \leq M \in \mathcal{O}(\text{poly}(n))$ and $\mu(\mathbf{x})$ being the average of $q_\theta(\mathbf{x})$ over $\boldsymbol{\theta}$. As $\max_{\mathbf{x} \in \mathcal{P}} [\mu(\mathbf{x})] \in \mathcal{O}(1/b^n)$ for some $b > 1$ (due to the assumption that the fixed points are exponentially small), this leads to $\mu_s \in \mathcal{O}(\text{poly}(n)/b^n)$.

Then, the variance can be upper bounded as

$$\sigma_s^2 = \text{Var}_{\boldsymbol{\theta}} \left[\sum_{\mathbf{x} \in \mathcal{P}} q_{\boldsymbol{\theta}}(\mathbf{x}) \right] \quad (21)$$

$$= \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x})] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \in \mathcal{P} \\ \mathbf{x} \neq \mathbf{x}'}} \text{Cov}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\mathbf{x}')] \quad (22)$$

$$\leq \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x})] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \in \mathcal{P} \\ \mathbf{x} \neq \mathbf{x}'}} \sqrt{\text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x})] \text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x}')] } \quad (23)$$

$$\leq \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x})] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \in \mathcal{P} \\ \mathbf{x} \neq \mathbf{x}'}} \frac{\text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x})] + \text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x}')] }{2} \quad (24)$$

$$= N_p \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x})] \quad (25)$$

$$\leq N_p^2 \max_{\mathbf{x} \in \mathcal{P}} [\sigma^2(\mathbf{x})], \quad (26)$$

where in the first inequality we have used Cauchy-Schwarz, and the second inequality is the inequality of arithmetic and geometric means $\sqrt{xy} \leq (x + y)/2$ for $x, y > 0$. $\sigma^2(\mathbf{x})$ is the variance of $q_{\boldsymbol{\theta}}(\mathbf{x})$ over $\boldsymbol{\theta}$. Finally, assuming that $q_{\boldsymbol{\theta}}(\mathbf{x}')$ exponentially concentrates over all bitstrings in the dataset, we have $\sigma_s \in \mathcal{O}(\text{poly}(n)/b^{b'})$ for some $b' > 1$.

Now, we are ready to continue from Eq. (15).

$$\Pr_{\tilde{Q}_{\boldsymbol{\theta}}, \boldsymbol{\theta}} [\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\boldsymbol{\theta}})] \geq (1 - (\mu_s + \sqrt{\sigma_s}))^N (1 - \sigma_s) \quad (27)$$

$$\geq (1 - N(\mu_s + \sqrt{\sigma_s}))(1 - \sigma_s) \quad (28)$$

$$\geq 1 - (N(\mu_s + \sqrt{\sigma_s}) + \sigma_s) \quad (29)$$

$$\geq 1 - \left(N \left(N_p \max_{\mathbf{x} \in \mathcal{P}} [\mu(\mathbf{x})] + \sqrt{N_p \max_{\mathbf{x} \in \mathcal{P}} [\sigma(\mathbf{x})]} \right) + N_p \max_{\mathbf{x} \in \mathcal{P}} [\sigma(\mathbf{x})] \right) \quad (30)$$

$$= 1 - \delta, \quad (31)$$

where in the second inequality we use Bernoulli's inequality, the third inequality is due to dropping the positive term when expanding the product, and in the last inequality we use Eq. (20) and Eq. (26). In the last line, we denote $\delta = N \left(N_p \max_{\mathbf{x} \in \mathcal{P}} [\mu(\mathbf{x})] + \sqrt{N_p \max_{\mathbf{x} \in \mathcal{P}} [\sigma(\mathbf{x})]} \right) + N_p \max_{\mathbf{x} \in \mathcal{P}} [\sigma(\mathbf{x})]$ and given the polynomial scaling of N and N_p as well as the exponential scaling of $\max_{\mathbf{x} \in \mathcal{P}} [\mu(\mathbf{x})]$ and $\max_{\mathbf{x} \in \mathcal{P}} [\sigma(\mathbf{x})]$, we have

$$\delta \in \mathcal{O} \left(\frac{\text{poly}(n)}{c^n} \right), \quad (32)$$

for some $c > 1$. This completes the proof of the theorem. $\square$

B. No overlap for a model without inductive bias and with polynomial support

Theorem 1 applies for any parameterized circuit such that the model probabilities are concentrated over the bitstrings in the training set. This can occur if the model probabilities are exponentially concentrated over *all* bitstrings in $\mathcal{X}$. This is typically true for any unstructured circuit, such as those given in Proposition 2. However, the requirement that the model probabilities are concentrated over the bitstrings in the training set (but not necessarily all bitstrings) is much weaker. Crucially, this happens even when the model has polynomial support on the space of bitstrings (see Fig. 3 in the main text). As an example, we show that if the model does not impose a strong inductive bias that aligns with the target distribution, it is generally unlikely that samples from the model have any overlap with the training dataset, which in turn results in the model becoming untrainable. Specifically, we prove the following Supplemental Proposition.

Supplemental Proposition 1 (Concentration of pairwise explicit losses for models lacking inductive bias). *Consider the scenario where the generative model has polynomial support which is chosen at random. The probability that these bitstrings are not in the training set $\tilde{P}$ is exponentially close to 1. That is, with probability*

$$\Pr_{\tilde{Q}_\theta, \theta}[\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)] \geq 1 - \delta', \quad \delta' \in \mathcal{O}\left(\frac{\text{poly}(n)}{c'^n}\right), \quad (33)$$

for some $c' > 1$, and $\mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)$ the statistical estimate of the loss as defined Eq. (8).

Proof. Denote $N_{\mathcal{P}}$ and $N_{\mathcal{Q}}$ as the number of non-zero probabilities in the training and model distributions, respectively, i.e., their support. Given that the model support is chosen randomly, there are $\binom{2^n - N_{\mathcal{P}}}{N_{\mathcal{Q}}}$ ways of picking non-zero model probabilities such that the supports do not overlap. Then, the probability that there is no overlap between the two distributions is

$$\Pr_{\tilde{Q}_\theta, \theta}[\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)] = \frac{\binom{2^n - N_{\mathcal{P}}}{N_{\mathcal{Q}}}}{\binom{2^n}{N_{\mathcal{Q}}}} \quad (34)$$

$$= \frac{(2^n - N_{\mathcal{Q}}) \times \dots \times (2^n - N_{\mathcal{Q}} - N_{\mathcal{P}} + 1)}{2^n \times \dots \times (2^n - N_{\mathcal{P}} + 1)} \quad (35)$$

$$= \prod_{k=0}^{N_{\mathcal{P}}-1} \frac{2^n - N_{\mathcal{Q}} - k}{2^n - k} \quad (36)$$

$$= \prod_{k=0}^{N_{\mathcal{P}}-1} \left(1 - \frac{N_{\mathcal{Q}}}{2^n - k}\right) \quad (37)$$

$$\geq \left(1 - \frac{N_{\mathcal{Q}}}{2^n - N_{\mathcal{P}} + 1}\right)^{N_{\mathcal{P}}} \quad (38)$$

$$\geq 1 - \frac{N_{\mathcal{Q}} N_{\mathcal{P}}}{2^n - N_{\mathcal{P}} + 1}, \quad (39)$$

where the first lower bound is due to taking the smallest term in the product and the last inequality is due to the Bernoulli's inequality $(1 - x)^n \geq 1 - nx$. For the polynomial supports $N_{\mathcal{Q}}, N_{\mathcal{P}} \in \mathcal{O}(\text{poly}(n))$, the lower bound becomes exponentially close to 1, which completes the proof. $\square$

This Supplemental Proposition highlights that the fundamental problem underlying exponential concentration is the miss-alignment of the model probabilities and the training data. Any randomly chosen quantum circuit ansatz with random parametrization is bound to fail with an explicit loss because of the exponentially large space of bitstrings.

A concrete example that falls short is the case of near-identity initialization of the quantum circuit $U(\theta)$ of a QCBM, which corresponds to a near-zero initialization of the parameter vector θ . While in the context of VQA-type problems it has been shown to mitigate vanishing gradients at the initial training step [1], this strategy induces only significant probabilities on a polynomial number of bitstrings and thus leads to exponential concentration for general datasets. The reason is that the all-zero bitstring and bitstrings that are few bit-flips away from it have no *a priori* reason to be relevant to the modelling task.

A minimal expansion of the near-identity initialization, which appears to introduce inductive bias but does not necessarily so, is to initialize the quantum circuit model in one of the training states. That is, one sets $|\psi(\theta)\rangle = |\mathbf{x}_0\rangle$ in an attempt to avoid loss concentration around the initialization. While this does in fact exhibit initial gradients towards $q_\theta(\mathbf{x}_0) = \tilde{p}(\mathbf{x}_0)$, the chance that the model then contains a non-vanishing probability on a significant number of other samples $\mathbf{x} \neq \mathbf{x}_0$ is still low (or likely exponentially low). A possible exception would be a dataset exhibiting cluster behaviour in the space of bitstrings. Then one could indeed initialize in and around the centroid bitstring and likely achieve improved performance across various metrics.

C. Untrainability of the pairwise explicit loss function: Proof of Corollary 3

In this sub-section, we provide the proof of Corollary 3, which shows that the concentration of statistical estimate of the pairwise explicit loss makes such losses untrainable.

Corollary 3 (Untrainability of the pairwise explicit loss function). *Under the same conditions as in Theorem 1, the probability that the difference between the two statistical estimates of the loss function at θ_1 and θ_2 does not contain any information about the training distribution is exponentially close to 1. Particularly, we have*

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) - \tilde{\mathcal{L}}(\theta_2) = \Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2})] \geq 1 - 2\delta, \quad (40)$$

with $\delta \in \mathcal{O}\left(\frac{\text{poly}(n)}{c^n}\right)$ for some $c > 1$, $\tilde{Q}_{\theta_1}$ (and $\tilde{Q}_{\theta_2}$) is a set of sampling bitstrings obtained from the quantum generative model at the parameter value θ_1 (and θ_2), as well as

$$\Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2}) = \sum_{\mathbf{x} \in \mathcal{Q}_{\theta_1}} f(0, \tilde{q}_{\theta_1}(\mathbf{x})) - \sum_{\mathbf{x} \in \mathcal{Q}_{\theta_2}} f(0, \tilde{q}_{\theta_2}(\mathbf{x})), \quad (41)$$

with $\mathcal{Q}_{\theta_1}$ (and $\mathcal{Q}_{\theta_2}$) being a set of unique bitstrings in $\tilde{Q}_{\theta_1}$ (and $\tilde{Q}_{\theta_2}$). Crucially, $\Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2})$ does not depend on any $\tilde{p}(\mathbf{x}) \in \tilde{P}$.

Proof. We first note that

$$\mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_1}) - \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_2}) = \left(\sum_{\mathbf{x} \in \tilde{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \tilde{Q}_{\theta_1}} f(0, \tilde{q}_{\theta_1}(\mathbf{x})) \right) - \left(\sum_{\mathbf{x} \in \tilde{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \tilde{Q}_{\theta_2}} f(0, \tilde{q}_{\theta_2}(\mathbf{x})) \right) \quad (42)$$

$$= \sum_{\mathbf{x} \in \tilde{Q}_{\theta_1}} f(0, \tilde{q}_{\theta_1}(\mathbf{x})) - \sum_{\mathbf{x} \in \tilde{Q}_{\theta_2}} f(0, \tilde{q}_{\theta_2}(\mathbf{x})) \quad (43)$$

$$= \Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2}). \quad (44)$$

As estimating the loss function at θ_1 and θ_2 are two independent events for $|\tilde{Q}_{\theta_1}|, |\tilde{Q}_{\theta_2}| \in \mathcal{O}(\text{poly}(n))$, we have

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) - \tilde{\mathcal{L}}(\theta_2) = \Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2})] = \Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_1}) \cap \tilde{\mathcal{L}}(\theta_2) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_2})] \quad (45)$$

$$= \Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_1})] \cdot \Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_2) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_2})] \quad (46)$$

$$\geq (1 - \delta)(1 - \delta) \quad (47)$$

$$\geq 1 - 2\delta, \quad (48)$$

where the second equality is due to the independence of two events, the first inequality is from applying Theorem 1 and the last line is from dropping the δ^2 term. $\square$

III. SUPPLEMENTARY NOTE - EXACT VARIANCE OF KLD – LARGE VARIANCE IS NOT EVERYTHING

In this section, we show that the variance of the exact KL-divergence can be polynomially large even when model probabilities exhibit exponential concentration. Instead, the scaling of the variance depends on the number of supports of the training/target distribution. This is summarised by Proposition 1 in the main text which we repeat here for convenience.

Proposition 1. *Consider the KLD loss as defined in Eq. (4) in the main text. Assume access to the exact target distribution $p(\mathbf{x})$ and the model distribution $\tilde{q}(\mathbf{x})$. Then, we have*

- For deep (Haar random) parametrized circuit $U(\boldsymbol{\theta})$, the variance of the loss scales asymptotically ($2^n \gg 1$) as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}^{\text{KLD}}(\boldsymbol{\theta})] = \frac{\pi^2}{6} \sum_{\mathbf{x}} p^2(\mathbf{x}) . \quad (49)$$

- For a random tensor product circuit $U(\boldsymbol{\theta}) = \bigotimes_{i=1}^n U_i(\theta_i)$ where $U_i(\theta_i)$ is a random single-qubit unitary, the variance of the loss scales asymptotically as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}^{\text{KLD}}(\boldsymbol{\theta})] = n - \frac{\pi^2}{6} \sum_{\mathbf{x}, \mathbf{x}'} p(\mathbf{x}) p(\mathbf{x}') \|\mathbf{x} - \mathbf{x}'\|_H , \quad (50)$$

where $\|\cdot\|_H$ is the Hamming distance.

Proof. The variance of KLD loss can be expressed as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}^{\text{KLD}}(\boldsymbol{\theta})] = \text{Var}_{\boldsymbol{\theta}} \left[\sum_{\mathbf{x}} p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q_{\boldsymbol{\theta}}(\mathbf{x})} \right) \right] \quad (51)$$

$$= \text{Var}_{\boldsymbol{\theta}} \left[\sum_{\mathbf{x}} p(\mathbf{x}) \log (q_{\boldsymbol{\theta}}(\mathbf{x})) \right] \quad (52)$$

$$= \sum_{\mathbf{x}} p^2(\mathbf{x}) \text{Var}_{\boldsymbol{\theta}}[\log (q_{\boldsymbol{\theta}}(\mathbf{x}))] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \\ \mathbf{x} \neq \mathbf{x}'}} p(\mathbf{x}) p(\mathbf{x}') \text{Cov}_{\boldsymbol{\theta}} [\log (q_{\boldsymbol{\theta}}(\mathbf{x})), \log (q_{\boldsymbol{\theta}}(\mathbf{x}'))] , \quad (53)$$

where the second equality is from $\text{Var}_{\alpha}[aA(\alpha) + b] = a^2 \text{Var}_{\alpha}[A(\alpha)]$ with some constants a, b and in the third equality we use the identity $\text{Var}_{\alpha}[\sum_i A_i(\alpha)] = \sum_{i,j} \text{Cov}_{\alpha}[A_i(\alpha), A_j(\alpha)] = \sum_i \text{Var}_{\alpha}[A_i(\alpha)] + \sum_{i \neq j} \text{Cov}_{\alpha}[A_i(\alpha), A_j(\alpha)]$.

To compute each term in Eq. (53), we consider two different types of parametrized circuits: (i) the Haar random parametrized circuit and (ii) the tensor product circuit.

(i) Haar random parametrized circuit. Consider an ensemble of parametrized unitaries $\mathcal{U}_{\boldsymbol{\theta}} = \{U(\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ that forms a Haar random ensemble. In this scenario with $N_{\text{dim}} = 2^n \gg 1$, the distribution of output probabilities $q_{\boldsymbol{\theta}}(\mathbf{x})$ follows the so-called Porter-Thomas distribution $\Pr(q := q_{\boldsymbol{\theta}}(\mathbf{x})) = N_{\text{dim}} e^{-N_{\text{dim}} q}$ [2]. One can see that distribution is not exactly normalized $\int_0^1 \Pr(q) dq = \int_0^1 N_{\text{dim}} e^{-N_{\text{dim}} q} dq = 1 - e^{-N_{\text{dim}}}$ and has a correction term that is exponentially small in the dimension of the Hilbert space (i.e., doubly exponential in the number of qubits n).

In this regime, we can perform integration over the Porter-Thomas distribution instead of over the parameter distribution to compute an average associated with the model probability. That is, for some function $f(\cdot)$, we have

$$\mathbb{E}_{\boldsymbol{\theta}}[f(q_{\boldsymbol{\theta}}(\mathbf{x}))] = \int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \Pr(\boldsymbol{\theta}) f(q_{\boldsymbol{\theta}}(\mathbf{x})) = \int_0^1 dq \Pr(q) f(q) . \quad (54)$$

Now, consider the first term in Eq. (53). The average of $\log (q_{\boldsymbol{\theta}}(\mathbf{x}))$ can be written as (with $u = N_{\text{dim}} q$)

$$\mathbb{E}_{\boldsymbol{\theta}}[\log (q_{\boldsymbol{\theta}}(\mathbf{x}))] = \int_0^1 dq N_{\text{dim}} e^{-N_{\text{dim}} q} \log (q) \quad (55)$$

$$= \int_0^{N_{\text{dim}}} e^{-u} \log (u) du - \log (N_{\text{dim}}) \int_0^{N_{\text{dim}}} e^{-u} du \quad (56)$$

$$= \int_0^{\infty} e^{-u} \log (u) du - \int_{N_{\text{dim}}}^{\infty} e^{-u} \log (u) du - \log (N_{\text{dim}}) (1 - e^{-N_{\text{dim}}}) \quad (57)$$

$$= \int_0^{\infty} e^{-u} \log (u) du - \left(-e^{-u} \log (u) \right) \Big|_{u=N_{\text{dim}}}^{\infty} + \int_{N_{\text{dim}}}^{\infty} \frac{e^{-u}}{u} du - \log (N_{\text{dim}}) (1 - e^{-N_{\text{dim}}}) \quad (58)$$

$$= -(\gamma + \log (N_{\text{dim}}) + E_1(N_{\text{dim}})) , \quad (59)$$

where $\gamma = -\int_0^\infty e^{-u} \log(u) du$ is the Euler Mascheroni constant and $E_1(N_{\text{dim}}) = \int_{N_{\text{dim}}}^\infty \frac{e^{-u}}{u} du$ is the exponential integral. In the above derivations, the second equality is by the change of variable $u = N_{\text{dim}} q$, the fourth equality is due to partial integration on the middle term.

Similarly, the second moment of $\log(q_\theta(\mathbf{x}))$ is of the form (again, with $u = N_{\text{dim}} q$)

$$\mathbb{E}_\theta[\log^2(q_\theta(\mathbf{x}))] = \int_0^1 dq N_{\text{dim}} e^{-N_{\text{dim}} q} \log^2(q) \quad (60)$$

$$= \int_0^{N_{\text{dim}}} e^{-u} \log^2(u) du - 2 \log(N_{\text{dim}}) \int_0^{N_{\text{dim}}} e^{-u} \log(u) du + \log^2(N_{\text{dim}}) \int_0^{N_{\text{dim}}} e^{-u} du \quad (61)$$

$$= \int_0^\infty e^{-u} \log^2(u) du - \int_{N_{\text{dim}}}^\infty e^{-u} \log^2(u) du - 2 \log(N_{\text{dim}}) \int_0^{N_{\text{dim}}} e^{-u} \log(u) du + \log^2(N_{\text{dim}})(1 - e^{-N_{\text{dim}}}) \quad (62)$$

$$= \gamma^2 + \frac{\pi^2}{6} - \int_{N_{\text{dim}}}^\infty e^{-u} \log^2(u) du - 2 \log(N_{\text{dim}}) \int_0^{N_{\text{dim}}} e^{-u} \log(u) du + \log^2(N_{\text{dim}})(1 - e^{-N_{\text{dim}}}), \quad (63)$$

where $\int_0^\infty e^{-u} \log^2(u) du = \gamma^2 + \frac{\pi^2}{6}$ as a result of the direct computation and $\int_0^{N_{\text{dim}}} e^{-u} \log(u) du$ can be handled in an identical way to the average shown above. For $N_{\text{dim}} \gg 1$, we have $\int_{N_{\text{dim}}}^\infty e^{-u} \log^2(u) du \approx e^{-N_{\text{dim}}} \log^2(N_{\text{dim}})$ which can be verified using the L'Hôpital's rule

$$\lim_{N_{\text{dim}} \rightarrow \infty} \frac{e^{-N_{\text{dim}}} \log^2 N_{\text{dim}}}{\int_{N_{\text{dim}}}^\infty e^{-u} \log^2(u) du} = \lim_{N_{\text{dim}} \rightarrow \infty} \frac{-e^{-N_{\text{dim}}} \log^2 N_{\text{dim}} + e^{-N_{\text{dim}}} \frac{2 \log N_{\text{dim}}}{N_{\text{dim}}}}{-e^{-N_{\text{dim}}} \log^2(N_{\text{dim}})} = \lim_{N_{\text{dim}} \rightarrow \infty} 1 - \frac{2}{N_{\text{dim}} \log N_{\text{dim}}} = 1. \quad (64)$$

Additionally, in this regime the exponential integral approaches (e.g., see Ref. [3])

$$\lim_{N_{\text{dim}} \rightarrow \infty} E_1(N_{\text{dim}}) = \lim_{N_{\text{dim}} \rightarrow \infty} \int_{N_{\text{dim}}}^\infty \frac{e^{-u}}{u} du \approx \frac{e^{-N_{\text{dim}}}}{N_{\text{dim}}}. \quad (65)$$

Hence, for $N_{\text{dim}} \gg 1$, the variance of $\log(q_\theta(\mathbf{x}))$ scales as

$$\text{Var}_\theta[\log(q_\theta(\mathbf{x}))] = \frac{\pi^2}{6}. \quad (66)$$

Next in the menu, we show that the covariance terms in Eq. (53) disappear. In other words, $q_\theta(\mathbf{x})$ and $q_\theta(\mathbf{x}')$ are independent and their joint distribution respects the product formulae. To see this, we follow a similar approach in Ref. [2] which derives the Porter Thomas distribution for a random quantum state. We write a quantum state in the computational basis as $|\psi\rangle = \sum_{\mathbf{x}} (a_{\mathbf{x}} + i b_{\mathbf{x}}) |\mathbf{x}\rangle$. When $|\psi\rangle$ is uniformly chosen at random from the Hilbert space, there is only the normalization constrain on the probabilities i.e., $\sum_{\mathbf{x}} q(\mathbf{x}) = \sum_{\mathbf{x}} (a_{\mathbf{x}}^2 + b_{\mathbf{x}}^2) = 1$. The joint distribution of probabilities q and q' can be derived by a ratio of the volume of the Hilbert space corresponding to states with q and q' , $\mathcal{V}_{q,q'}$ to the total volume of all states, $\mathcal{V}_{\text{tot}}$. More precisely, we can use delta functions to select the relevant states and compute the associated volumes and the joint distribution as

$$\text{Pr}(q, q') = \frac{\mathcal{V}_{q,q'}}{\mathcal{V}_{\text{tot}}}, \quad (67)$$

with

$$\mathcal{V}_{q,q'} = \int_{-\infty}^\infty \prod_{\mathbf{x}} da_{\mathbf{x}} db_{\mathbf{x}} \delta\left(1 - \sum_{\mathbf{x}} (a_{\mathbf{x}}^2 + b_{\mathbf{x}}^2)\right) \delta(a_{\mathbf{x}_0}^2 + b_{\mathbf{x}_0}^2 - q) \delta(a_{\mathbf{x}_1}^2 + b_{\mathbf{x}_1}^2 - q'), \quad (68)$$

$$\mathcal{V}_{\text{tot}} = \int_{-\infty}^\infty \prod_{\mathbf{x}} da_{\mathbf{x}} db_{\mathbf{x}} \delta\left(1 - \sum_{\mathbf{x}} (a_{\mathbf{x}}^2 + b_{\mathbf{x}}^2)\right), \quad (69)$$

where $\delta(1 - \sum_{\mathbf{x}}(a_{\mathbf{x}}^2 + b_{\mathbf{x}}^2))$ enforces the normalization condition, and $\delta(a_{\mathbf{x}_0}^2 + b_{\mathbf{x}_0}^2 - q)$ as well $\delta(a_{\mathbf{x}_1}^2 + b_{\mathbf{x}_1}^2 - q')$ correspond to select states that include q and q' probabilities, respectively. By expressing a delta function as $\delta(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt e^{i\omega t}$, the numerator can be expressed as

$$\mathcal{V}_{q,q'} = \int_{-\infty}^{\infty} \prod_{\mathbf{x}} da_{\mathbf{x}} db_{\mathbf{x}} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} dt e^{it(\sum_{\mathbf{x}}(a_{\mathbf{x}}^2 + b_{\mathbf{x}}^2) - 1)} \right) \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} ds e^{is(a_{\mathbf{x}_0}^2 + b_{\mathbf{x}_0}^2 - q)} \right) \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} dw e^{iw(a_{\mathbf{x}_1}^2 + b_{\mathbf{x}_1}^2 - q')} \right) \quad (70)$$

$$= \frac{1}{(2\pi)^3} \left[\int_{-\infty}^{\infty} dt e^{-it} \left(\int_{-\infty}^{\infty} da_{\mathbf{x}} db_{\mathbf{x}} e^{it(a_{\mathbf{x}}^2 + b_{\mathbf{x}}^2)} \right)^{N_{\text{dim}}-2} \right] \cdot \left[\int_{-\infty}^{\infty} ds e^{-isq} \int_{-\infty}^{\infty} da_{\mathbf{x}_0} db_{\mathbf{x}_0} e^{i(s+t)(a_{\mathbf{x}_0}^2 + b_{\mathbf{x}_0}^2)} \right] \cdot \left[\int_{-\infty}^{\infty} dw e^{-iwq'} \int_{-\infty}^{\infty} da_{\mathbf{x}_1} db_{\mathbf{x}_1} e^{i(w+t)(a_{\mathbf{x}_1}^2 + b_{\mathbf{x}_1}^2)} \right] \quad (71)$$

$$= \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} dt e^{-it} \left(\frac{i\pi}{t} \right)^{N_{\text{dim}}-2} \int_{-\infty}^{\infty} ds e^{-isq} \left(\frac{i\pi}{t+s} \right) \int_{-\infty}^{\infty} dw e^{-iwq'} \left(\frac{i\pi}{t+w} \right) \quad (72)$$

$$= \frac{(i\pi)^{N_{\text{dim}}}}{(2\pi)^3} (-2\pi i)^2 \int_{-\infty}^{\infty} dt \frac{e^{-it(1-q-q')}}{t^{N_{\text{dim}}-2}} \quad (73)$$

$$= \frac{(i\pi)^{N_{\text{dim}}}}{(2\pi)^3} (-2\pi i)^3 \frac{(-i(1-q-q'))^{N_{\text{dim}}-3}}{(N_{\text{dim}}-3)!} \quad (74)$$

$$= \frac{\pi^{N_{\text{dim}}}(1-q-q')^{N_{\text{dim}}-3}}{(N_{\text{dim}}-3)!}, \quad (75)$$

where we recognise Gaussian integration in the third equality i.e., $\int_{-\infty}^{\infty} dz e^{i\alpha z^2} = \sqrt{\frac{i\pi}{\alpha}}$. In the fourth equality, we apply the Cauchy residue theorem for the simple pole to obtain $\int_{-\infty}^{\infty} dz \frac{e^{-izp}}{t+z} = -2\pi i e^{itp}$. In the fifth equality, the integration of a higher order pole can be computed by adding infinitesimally small imaginary offset

$$\int_{-\infty}^{\infty} dz \frac{e^{-iz}}{z^N} = \lim_{\epsilon \rightarrow 0} \int_{-\infty+i\epsilon}^{\infty+i\epsilon} dz \frac{e^{-iz}}{z^N} = \frac{-2\pi i}{(N-1)!} \left(\frac{d^N}{dz^N} e^{-iz} \right) \Big|_{z=0} = \frac{-2\pi i}{(N-1)!} \cdot (-i)^{N-1}, \quad (76)$$

We refer the readers to Ref. [2, 4] for more details of the complex integration.

Similarly, the total volume of all states is of the form

$$\mathcal{V}_{\text{tot}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt e^{-it} \left(\int_{-\infty}^{\infty} da_{\mathbf{x}} db_{\mathbf{x}} e^{it(a_{\mathbf{x}}^2 + b_{\mathbf{x}}^2)} \right)^{N_{\text{dim}}} \quad (77)$$

$$= \frac{\pi^{N_{\text{dim}}}}{(N_{\text{dim}}-1)!}. \quad (78)$$

For large $N_{\text{dim}} \gg 1$, the joint distribution of model probabilities can then be expressed as

$$\Pr(q, q') = (N_{\text{dim}}-1)(N_{\text{dim}}-2)(1-q-q')^{N_{\text{dim}}-3} \approx N_{\text{dim}}^2 e^{-N_{\text{dim}}(q+q')} = \Pr(q)\Pr(q'), \quad (79)$$

which leads to the vanishing of covariance terms in Eq. (53)

$$\text{Cov}_{\boldsymbol{\theta}} [\log(q_{\boldsymbol{\theta}}(\mathbf{x})), \log(q_{\boldsymbol{\theta}}(\mathbf{x}'))] = 0. \quad (80)$$

Together with Eq (66), the variance in Eq. (53) in the regime of large N_{dim} can be written as

$$\text{Var}_{\boldsymbol{\theta}} [\mathcal{L}^{\text{KLD}}(\boldsymbol{\theta})] = \frac{\pi^2}{6} \sum_{\mathbf{x}} p^2(\mathbf{x}). \quad (81)$$

We remind the reader again that this result is valid in the large N_{dim} limit and for the uncorrelated distributions the error we are making from this approximation scales at least polynomially in N_{dim} i.e. exponentially with n . This completes the proof of the first part.

(ii) *Tensor product circuit.* Consider the parametrized circuit of the form $U(\boldsymbol{\theta}) = \otimes_{i=1}^n U_i(\theta_i)$ where $U_i(\theta_i)$ is an arbitrary single-qubit unitary acting on qubit i . In this scenario, the model probability is a product of one-bit marginal probabilities $q_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{i=1}^n q_{\theta_i}(x_i)$ where $q_{\theta_i}(x_i) = \text{Tr}[U_i(\theta_i)\rho_0^{(i)}U_i^\dagger(\theta_i)|x_i\rangle\langle x_i|]$ with $\rho_0^{(i)}$ being a reduced initial state on the qubit i , leading to $\log(q_{\boldsymbol{\theta}}(\mathbf{x})) = \sum_{i=1}^n \log(q_{\theta_i}(x_i))$. In addition, these one-bit marginal probabilities are independent which result in

$$\text{Cov}_{\boldsymbol{\theta}}[\log(q_{\theta_i}(x_i)), q_{\theta_j}(x_j)] = 0, \quad (82)$$

for $i \neq j$. With these, the variance can be simplified as follows

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}^{\text{KLD}}(\boldsymbol{\theta})] = \sum_{\mathbf{x}} p^2(\mathbf{x}) \sum_{i=1}^n \text{Var}_{\theta_i}[\log(q_{\theta_i}(x_i))] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \\ \mathbf{x} \neq \mathbf{x}'}} p(\mathbf{x})p(\mathbf{x}') \sum_{i=1}^n \text{Cov}_{\theta_i}[\log(q_{\theta_i}(x_i)), \log(q_{\theta_i}(x'_i))] \quad (83)$$

$$= \sum_{\mathbf{x}, \mathbf{x}'} p(\mathbf{x})p(\mathbf{x}') \sum_{i=1}^n \text{Cov}_{\theta_i}[\log(q_{\theta_i}(x_i)), \log(q_{\theta_i}(x'_i))] . \quad (84)$$

Here, we consider a single-qubit unitary parametrized by Euler angles i.e., $\theta_i = (a_i, b_i, c_i) \in [0, 2\pi] \times [0, \pi] \times [0, 2\pi]$ of the form [5]

$$U(\theta_i) = \begin{pmatrix} e^{i\frac{a_i+c_i}{2}} \cos \frac{b_i}{2} & e^{i\frac{a_i-c_i}{2}} \sin \frac{b_i}{2} \\ -e^{-i\frac{a_i-c_i}{2}} \sin \frac{b_i}{2} & e^{-i\frac{a_i+c_i}{2}} \cos \frac{b_i}{2} \end{pmatrix} . \quad (85)$$

This results in the one-bit marginal probability for $x_i = 0, 1$ as

$$q_{\theta_i}(0) = \cos^2\left(\frac{b_i}{2}\right), \quad q_{\theta_i}(1) = \sin^2\left(\frac{b_i}{2}\right), \quad (86)$$

and the Haar measure in this case is $d\mu_2 := \text{Pr}(\theta_i)d\theta_i = \sin(b_i)db_ida_idc_i$. As $q_{\theta_i}(x_i)$ only depends on b_i , then we need to compute average over b_i such that for any function $f(b_i)$ the average is given by $\mathbb{E}_{b_i}[f(b_i)] = \frac{1}{2} \int_0^\pi db_i \sin(b_i)f(b_i)$ where the factor $1/2$ is to ensure normalisation (i.e. $\mathbb{E}_{b_i}[1] = 1$). We give the following integrals that are used to compute the KLD variance

$$\mathbb{E}_{\theta_i}[\log(q_{\theta_i}(0))] = \int_0^\pi dx \sin(x) \log(\cos(x/2)) = -1, \quad (87)$$

$$\mathbb{E}_{\theta_i}[\log(q_{\theta_i}(1))] = \int_0^\pi dx \sin(x) \log(\sin(x/2)) = -1, \quad (88)$$

$$\mathbb{E}_{\theta_i}[\log^2(q_{\theta_i}(0))] = 2 \int_0^\pi dx \sin(x) \log^2(\cos(x/2)) = 2, \quad (89)$$

$$\mathbb{E}_{\theta_i}[\log^2(q_{\theta_i}(1))] = 2 \int_0^\pi dx \sin(x) \log^2(\sin(x/2)) = 2, \quad (90)$$

$$\mathbb{E}_{\theta_i}[\log(q_{\theta_i}(0)) \log(q_{\theta_i}(1))] = 2 \int_0^\pi dx \sin(x) \log(\cos(x/2)) \log(\sin(x/2)) = 2 - \frac{\pi^2}{6}, \quad (91)$$

which in turn leads to

$$\text{Var}_{\theta_i}[\log(q_{\theta_i}(0))] = \text{Var}_{\theta_i}[\log(q_{\theta_i}(1))] = 1, \quad (92)$$

$$\text{Cov}_{\theta_i}[\log(q_{\theta_i}(0)), \log(q_{\theta_i}(1))] = 1 - \frac{\pi^2}{6} \approx -0.645. \quad (93)$$

Inserting this into (84) leads to

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}^{\text{KLD}}(\boldsymbol{\theta})] = \sum_{\mathbf{x}, \mathbf{x}'} p(\mathbf{x})p(\mathbf{x}') \sum_{i=1}^n \left(\delta_{x_i, x'_i} + (1 - \delta_{x_i, x'_i}) \left(1 - \frac{\pi^2}{6}\right) \right) \quad (94)$$

$$= \sum_{\mathbf{x}, \mathbf{x}'} p(\mathbf{x})p(\mathbf{x}') \sum_{i=1}^n \left(1 - \frac{\pi^2}{6} (1 - \delta_{x_i, x'_i}) \right) \quad (95)$$

$$= n - \frac{\pi^2}{6} \sum_{\mathbf{x}, \mathbf{x}'} p(\mathbf{x})p(\mathbf{x}') \sum_{i=1}^n |x_i - x'_i| \quad (96)$$

$$= n - \frac{\pi^2}{6} \sum_{\mathbf{x}, \mathbf{x}'} p(\mathbf{x})p(\mathbf{x}') \|\mathbf{x} - \mathbf{x}'\|_H, \quad (97)$$

$$(98)$$

where the third equality uses $|x_i - x'_i| = 1 - \delta_{x_i, x'_i}$ for $x_i, x'_i \in \{0, 1\}$ and in the fourth equality we introduce the hamming distance between two bit-string i.e. $\sum_{i=1}^n |x_i - x'_i| = \|\mathbf{x} - \mathbf{x}'\|_H$. This completes the proof of the proposition. $\square$

As a final comment on the significance of Proposition 1, we remark that it is likely that if $\log(q_{\boldsymbol{\theta}}(\mathbf{x}))$ could be measured directly from quantum computers (and not as a post process of samples obtained from quantum computers), this would lead to a model with substantial (at worst polynomially vanishing) loss variances. Hence, this suggests a new (if very ambitious!) research opportunity to circumvent barren plateaus. Lastly, even with the log function measured directly and/or a direct access to the model probabilities, one could still face untrainability induced by a target distribution. For a deep parametrized circuit this happens when the purity of the target distribution becomes exponentially small i.e., $\sum_{\mathbf{x}} p^2(\mathbf{x}) \in \mathcal{O}(b^{-n})$ for some $b > 1$. The last panel in Fig 2 in the main text numerically supports this phenomenon for the cardinality dataset which has exponentially many support for deep circuits. This presents a target-oriented barrier to trainability similar to Ref. [6].

IV. SUPPLEMENTARY NOTE - ANALYSIS ON THE MMD LOSS FUNCTION

In this section, we provide analysis on the MMD loss functions, including detailed proofs as well as further discussion of our analytical results in the main text. Specifically, Supplementary Note IV A shows how the MMD loss can be viewed as the expectation of an observable and analyzes how its properties depend on the bandwidth σ . A detailed analysis of the MMD loss landscape for a tensor product QCBM is provided in Supplementary Note IV B. The trainability results are extended to a general Pauli rotation QCBM in Supplementary Note IV C. Lastly, in Supplementary Note IV D, we investigate the suitability of the MMD for learning global properties of a target distribution depending on our choice in bandwidth parameter.

For convenience, we start by recalling that the MMD loss is of the form

$$\mathcal{L}_{\text{MMD}}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})K(\mathbf{x}, \mathbf{y}) - 2 \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}) + \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} p(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}), \quad (99)$$

with the classical Gaussian kernel

$$K_{\sigma}(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma}} \quad (100)$$

$$= \prod_{i=1}^n e^{-\frac{(x_i - y_i)^2}{2\sigma}}, \quad (101)$$

where $\|\cdot\|_2$ is the 2-norm, $\sigma > 0$ is the so-called *bandwidth* parameter, and x_i, y_i are the value of bit i in bitstrings $\mathbf{x}, \mathbf{y}$ (of length n), respectively.

A. MMD as an observable

In this section, we explain how the MMD loss function can be seen as an expectation value of some observable and analyse how the observable behaves for different values of the kernel bandwidth.

We start by noting that each term in the MMD loss function can be seen as the expectation value of an observable

$$\mathcal{M}(\rho, \rho') = \text{Tr} \left[O_{\text{MMD}}^{(\sigma)} (\rho \otimes \rho') \right], \quad (102)$$

with the MMD observable defined as

$$O_{\text{MMD}}^{(\sigma)} := \sum_{\mathbf{x}, \mathbf{y}} K_{\sigma}(\mathbf{x}, \mathbf{y}) |\mathbf{x}\rangle\langle\mathbf{x}| \otimes |\mathbf{y}\rangle\langle\mathbf{y}|, \quad (103)$$

which acts on $2n$ qubits. To obtain each term in the MMD, ρ and ρ' can be either the quantum state of our QCBM model $\rho_{\theta} = |\psi(\theta)\rangle\langle\psi(\theta)|$ and/or the quantum state corresponding to the training data $\rho_{\tilde{p}}$ such that $\tilde{p}(\mathbf{x}) = \text{Tr}[\rho_{\tilde{p}}|\mathbf{x}\rangle\langle\mathbf{x}|]$. In particular, for computing the first MMD term, we have both $\rho = \rho' = \rho_{\theta}$ and, for computing the cross-term, we have $\rho' = \rho_{\tilde{p}}$ instead, and for the final term $\rho = \rho' = \rho_{\tilde{p}}$.

The MMD observable $O_{\text{MMD}}^{(\sigma)}$ can be rewritten in Pauli basis using $|\mathbf{x}\rangle\langle\mathbf{x}| = \bigotimes_{i=1}^n |x_i\rangle\langle x_i| = \bigotimes_{i=1}^n \frac{1}{2}(\mathbb{1}_i + (-1)^{x_i} Z_i)$ for the first n qubits and $|\mathbf{y}\rangle\langle\mathbf{y}| = \bigotimes_{i=1}^n |y_i\rangle\langle y_i| = \bigotimes_{i=1}^n \frac{1}{2}(\mathbb{1}_{n+i} + (-1)^{y_i} Z_{n+i})$ for the last n qubits, leading to

$$O_{\text{MMD}}^{(\sigma)} = \sum_{\mathbf{x}, \mathbf{y}} \bigotimes_{i=1}^n \left[\left(\frac{\mathbb{1}_i + (-1)^{x_i} Z_i}{2} \right) \otimes \left(\frac{\mathbb{1}_{n+i} + (-1)^{y_i} Z_{n+i}}{2} \right) \exp \left(-\frac{(x_i - y_i)^2}{2\sigma} \right) \right] \quad (104)$$

$$= \bigotimes_{i=1}^n \sum_{x_i, y_i} \left[\left(\frac{\mathbb{1}_i + (-1)^{x_i} Z_i}{2} \right) \otimes \left(\frac{\mathbb{1}_{n+i} + (-1)^{y_i} Z_{n+i}}{2} \right) \exp \left(-\frac{(x_i - y_i)^2}{2\sigma} \right) \right] \quad (105)$$

$$= \bigotimes_{i=1}^n [(1 - p_{\sigma}) \mathbb{1}_i \otimes \mathbb{1}_{n+i} + p_{\sigma} Z_i \otimes Z_{n+i}] \quad (106)$$

$$= \sum_{A \subseteq \mathcal{N}} (1 - p_{\sigma})^{n-|A|} p_{\sigma}^{|A|} \bigotimes_{i \in A} (Z_i \otimes Z_{n+i}) \quad (107)$$

$$= \sum_{l=0}^n \binom{n}{l} (1 - p_{\sigma})^{n-l} p_{\sigma}^l D_{2l}, \quad (108)$$

where we denote

$$p_{\sigma} = (1 - e^{-1/(2\sigma)})/2, \quad (109)$$

ranging between 0 and 1/2. The second equality is obtained using $\sum_{\mathbf{x}, \mathbf{y}} \bigotimes_{i=1}^n h(x_i, y_i) = \bigotimes_{i=1}^n \sum_{x_i, y_i} h(x_i, y_i)$ and the third one by explicitly summing over the x_i and y_i , each of which has two possible values of 0 and 1. To obtain the fourth inequality we expand the tensor product out explicitly and introduce the notation A to denote the set of all possible subsets of the indices $1, \dots, n$. That is,

$$A \subseteq \mathcal{N} = \{1, 2, \dots, n\}. \quad (110)$$

For example for $n = 3$, we have

$$A \in \{\{\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}. \quad (111)$$

In the last step, we denote

$$D_{2l} = \frac{1}{\binom{n}{l}} \sum_{\substack{A \subseteq \mathcal{N} \\ |A|=l}} \bigotimes_{i \in A} (Z_i \otimes Z_{n+i}), \quad (112)$$

which is a normalized sum of Pauli strings, each of which contains Pauli-Z operators of length $2l$, i.e., $2l$ -body interactions.

Interestingly, p_σ can be seen as the probability of assigning a pair of single-qubit Pauli-Z operators to a Pauli string. As a consequence, the coefficient $w_\sigma(l)$ follows a binomial distribution,

$$w_\sigma(l) = \binom{n}{l} (1-p_\sigma)^{n-l} p_\sigma^l, \quad (113)$$

and can be interpreted as the probability of having D_{2l} i.e., all possible Pauli-strings with $2l$ Pauli-Z operators. By adopting this Monte Carlo sampling perspective, the MMD observable $O_{\text{MMD}}^{(\sigma)}$ can be constructed as a sampling average where the operator D_{2l} is sampled with the probability $w_\sigma(l)$. This allows us to analyze the dominant terms in $O_{\text{MMD}}^{(\sigma)}$ by using standard properties of the binomial distribution. For example, the largest $w_\sigma(l)$ (i.e. the mode of the distribution) occurs for

$$(n+1)p_\sigma - 1 \leq l_{\max} = \arg \max(w_\sigma(l)) \leq (n+1)p_\sigma. \quad (114)$$

We note that $w_\sigma(l)$ is monotonically increasing for $l < l_{\max}$ and is monotonically decreasing for $l > l_{\max}$. The average (i.e. mean) bodyness of the MMD observable is given by

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] = 2np_\sigma, \quad (115)$$

and its variance is

$$\text{Var}_{l \sim w_\sigma(l)}[2l] = 4np_\sigma(1-p_\sigma). \quad (116)$$

We are now ready to investigate how $O_{\text{MMD}}^{(\sigma)}$ depends on σ .

(i) For a constant bandwidth $\sigma \in \mathcal{O}(1)$: We have the following proposition.

Proposition 3 (MMD is global with a constant bandwidth). *For $\sigma \in \mathcal{O}(1)$, the average bodyness of the MMD operator containing Pauli terms with weight $w_\sigma(l)$ is*

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] \in \Theta(n). \quad (117)$$

Similarly, the variance in the bodyness is given by

$$\text{Var}_{l \sim w_\sigma(l)}[2l] \in \Theta(n). \quad (118)$$

Proof. By considering that $p_\sigma, (1-p_\sigma) \in \mathcal{O}(1)$ for $\sigma \in \mathcal{O}(1)$ in conjunction with Eqs. (115) and (116), we have the scaling of the average bodyness and its variance as claimed. $\square$

Together, this implies that, on average, we are likely to sample D_{2l} with $l \in \mathcal{O}(n)$. Now, we show that the contribution from low-body terms is negligible in this regime. Consider the sum of probabilities up to k bodies (such that $k < l_{\max}$)

$$\sum_{l=0}^k w_\sigma(l) = \sum_{l=0}^k \binom{n}{l} (1-p_\sigma)^{n-l} p_\sigma^l \quad (119)$$

$$\leq (k+1) \binom{n}{k} (1-p_\sigma)^{n-k} p_\sigma^k \quad (120)$$

$$\leq (k+1) \left(\frac{ne}{k}\right)^k (1-p_\sigma)^{n-k} p_\sigma^k, \quad (121)$$

where, in the first inequality we take the maximum value of the sum as each term in the sum is monotonically increasing before $l_{\max} \in \mathcal{O}(n)$, the second inequality is due to $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$. It is straight forward to see that, for $k \in \mathcal{O}(1)$,

$$\sum_{l=0}^k w_\sigma(l) \in \mathcal{O}(\text{poly}(n)/b^n), \quad (122)$$

for some $b > 0$. Altogether, for $\sigma \in \mathcal{O}(1)$, the MMD observable is global and the contribution from low-body interactions is exponentially suppressed in the number of qubits.

(ii) For a linearly-scaled bandwidth $\sigma \in \Theta(n)$: The situation here is opposite to what we have in the case (i). First we note that

$$p(\sigma) = \frac{1 - \left(1 - \frac{1}{2\sigma} + \mathcal{O}\left(\frac{1}{\sigma^2}\right)\right)}{2} \quad (123)$$

$$= \frac{1}{4\sigma} + \mathcal{O}\left(\frac{1}{\sigma^2}\right) \quad (124)$$

$$\in \mathcal{O}\left(\frac{1}{n}\right) \quad (125)$$

with $\sigma \in \Theta(n)$. Thus in this limit the average and variance of the bodyness of the MMD operator are given by

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] \in \mathcal{O}(1) \quad (126)$$

$$\text{Var}_{l \sim w_\sigma(l)}[2l] \in \mathcal{O}(1) . \quad (127)$$

Intuitively, this implies the MMD observable is largely composed of low-body contributions in this bandwidth regime. As a consequence, when computing the MMD loss, the contribution from global terms are negligible. This notion is formalized in Proposition 4 in the main text, which is proven below.

Proposition 4 (MMD consists largely of low-body terms for $\sigma \in \Theta(n)$). *Let $\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\boldsymbol{\theta})$ be a truncated MMD loss with a truncated operator $\tilde{O}_{\text{MMD}}^{(\sigma,k)}$ that contains up to the $2k$ -body interactions in $O_{\text{MMD}}^{(\sigma)}$,*

$$\tilde{O}_{\text{MMD}}^{(\sigma,k)} := \sum_{l=0}^k w_\sigma(l) D_{2l} , \quad (128)$$

where $w_\sigma(l)$ are Bernoulli-distributed weights defined in Eq. (113). For $\sigma \in \Theta(n)$, the difference between the exact and local approximation of the loss is bounded as

$$|\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\boldsymbol{\theta})| \leq \epsilon(k) , \quad (129)$$

with

$$\epsilon(k) \in \mathcal{O}\left(n(c/k)^k\right) , \quad (130)$$

for some positive constant c .

Proof. Let $\rho_{\boldsymbol{\theta}} = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|$ be the quantum state of our QCBM and $\rho_{\tilde{p}}$ with $\tilde{p}(\mathbf{x}) = \text{Tr}[\rho_{\tilde{p}}|\mathbf{x}\rangle\langle\mathbf{x}|]$ be the quantum

associated with the training data. We then have

$$|\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\boldsymbol{\theta})| = \left| \text{Tr} \left[\left(O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\boldsymbol{\theta}} \otimes \rho_{\boldsymbol{\theta}}) \right] - 2 \text{Tr} \left[\left(O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\boldsymbol{\theta}} \otimes \rho_{\bar{p}}) \right] \right. \\ \left. + \text{Tr} \left[\left(O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\bar{p}} \otimes \rho_{\bar{p}}) \right] \right| \quad (131)$$

$$\leq \left| \text{Tr} \left[\left(O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\boldsymbol{\theta}} \otimes \rho_{\boldsymbol{\theta}}) \right] \right| + 2 \left| \text{Tr} \left[\left(O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\boldsymbol{\theta}} \otimes \rho_{\bar{p}}) \right] \right| \\ + \left| \text{Tr} \left[\left(O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\bar{p}} \otimes \rho_{\bar{p}}) \right] \right| \quad (132)$$

$$\leq \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \left\| \rho_{\boldsymbol{\theta}} \otimes \rho_{\boldsymbol{\theta}} \right\|_1 + 2 \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \left\| \rho_{\boldsymbol{\theta}} \otimes \rho_{\bar{p}} \right\|_1 \\ + \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \left\| \rho_{\bar{p}} \otimes \rho_{\bar{p}} \right\|_1 \quad (133)$$

$$= 4 \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \quad (134)$$

$$= 4 \left\| \sum_{l=k+1}^n w_{\sigma}(l) D_{2l} \right\|_{\infty} \quad (135)$$

$$\leq 4 \sum_{l=k+1}^n \binom{n}{l} (1 - p_{\sigma})^{n-l} p_{\sigma}^l \quad (136)$$

$$\leq 4 \sum_{l=k+1}^n \binom{n}{l} \left(\frac{1 - e^{-1/2\sigma}}{2} \right)^l \quad (137)$$

$$\leq 4 \sum_{l=k+1}^n \left(\frac{ne}{4l\sigma} \right)^l, \quad (138)$$

where the first inequality is due to the triangle inequality, the second inequality is due to Hölder's inequality, the second equality is that the 1-norm of the quantum state is 1 (density operators have trace 1), the third inequality uses triangle inequality and the fact that the infinity norm of Pauli operators is 1, the fourth inequality is from $1 - p_{\sigma} \leq 1$, and in the last inequality we use $e^{-x} \geq 1 - x$ together with $\binom{n}{l} \leq \left(\frac{ne}{l} \right)^l$.

To further upper bound the truncation error consider $f(x) = \left(\frac{ne}{4\sigma x} \right)^x$ for $x > 0$. We notice that $f'(x) = f(x) [\ln \left(\frac{ne}{4\sigma x} \right) - 1]$ which leads to the maximum of $f(x)$ at $x^* = n/(4\sigma)$. This leads to

$$\sum_{l=k+1}^n \left(\frac{ne}{4\sigma l} \right)^l \leq (n - k) \left(\frac{ne}{4\sigma k^*} \right)^{k^*}, \quad (139)$$

where $k^* = \max(k, n/4\sigma)$. Finally, if we assume that $\sigma \in \Theta(n)$ and if $k \geq n/(4\sigma)$, then we obtain

$$\epsilon(k) \in \mathcal{O} \left(n \left(\frac{c}{k} \right)^k \right) \quad (140)$$

where $c = \frac{ne}{4\sigma} \in \mathcal{O}(1)$. This completes the proof. $\square$

B. MMD variance for a tensor product ansatz

Here we analyse the scaling of the MMD loss variance for a tensor product ansatz. In particular, we derive Theorem 2 in the main text and provide further discussion.

The variance of the MMD can be computed as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}(\boldsymbol{\theta})] = \text{Var}_{\boldsymbol{\theta}} \left[\sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) - 2 \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) + \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} p(\mathbf{x}) p(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) \right] \quad (141)$$

$$= \text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})] + 4 \text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{p,q}(\boldsymbol{\theta})] - 4 \text{Cov}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta}), \mathcal{K}_{p,q}(\boldsymbol{\theta})], \quad (142)$$

where we have used $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$ and $\text{Var}[X + c] = \text{Var}[X]$ for any random variables X, Y and some constant c . We also introduce the shorthand notation of the first and second terms in the MMD loss as

$$\mathcal{K}_{q,q}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) , \quad (143)$$

and

$$\mathcal{K}_{p,q}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) . \quad (144)$$

Throughout this sub-section, we consider the QCBM that is comprised of the tensor product ansatz which is of the form

$$U(\boldsymbol{\theta}) = \bigotimes_{i=1}^n U_i(\boldsymbol{\theta}_i) , \quad (145)$$

with $U_i(\boldsymbol{\theta}_i)$ being a single-qubit random unitary acting on qubit i such that its ensemble of over $\boldsymbol{\theta}_i$ i.e., $\{U_i(\boldsymbol{\theta}_i)\}_{\boldsymbol{\theta}_i}$ forms the single-qubit Haar random ensemble. The model probability of measuring a bitstring $\mathbf{x}$ can be expressed as

$$q_{\boldsymbol{\theta}}(\mathbf{x}) = \text{Tr} [U(\boldsymbol{\theta}) |\mathbf{0}\rangle \langle \mathbf{0}| U^\dagger(\boldsymbol{\theta}) |\mathbf{x}\rangle \langle \mathbf{x}|] \quad (146)$$

$$= \prod_{i=1}^n \text{Tr} [U_i(\boldsymbol{\theta}_i) |0_i\rangle \langle 0_i| U_i^\dagger(\boldsymbol{\theta}_i) |x_i\rangle \langle x_i|] . \quad (147)$$

where we use $(A \otimes B)(C \otimes D) = AC \otimes BD$ and $\text{Tr}[A \otimes B] = \text{Tr}[A] \text{Tr}[B]$.

1. Preliminaries: Haar integration and Pauli operators

Crucially, as the rotation angles $\boldsymbol{\theta}_i$ are independent and $\{U_i(\boldsymbol{\theta}_i)\}_{\boldsymbol{\theta}_i}$ is a single-qubit Haar random ensemble, averaging over $\boldsymbol{\theta}_i$ is equivalent to averaging over the single-qubit Haar ensemble. Hence, we can invoke Haar integration to perform an average over randomly initialized parameters $\boldsymbol{\theta}_i$ on each individual qubit. As an example, consider the average of the probability $q_{\boldsymbol{\theta}}(\mathbf{x})$ over single qubit Haar random product states

$$\mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x})] = \int dU(\boldsymbol{\theta}) \text{Tr} [U(\boldsymbol{\theta}) |\mathbf{0}\rangle \langle \mathbf{0}| U^\dagger(\boldsymbol{\theta}) |\mathbf{x}\rangle \langle \mathbf{x}|] \quad (148)$$

$$= \prod_{i=1}^n \int dU_i(\boldsymbol{\theta}_i) \text{Tr} [U_i(\boldsymbol{\theta}_i) |0_i\rangle \langle 0_i| U_i^\dagger(\boldsymbol{\theta}_i) |x_i\rangle \langle x_i|] \quad (149)$$

$$= \prod_{i=1}^n \frac{\text{Tr}[|0_i\rangle \langle 0_i|] \text{Tr}[|x_i\rangle \langle x_i|]}{2} \quad (150)$$

$$= \frac{1}{2^n} , \quad (151)$$

where we used the Haar integral formula $\int dV V M V^\dagger = \text{Tr}[M]/d_V$ (with d_V as dimension of V). The Haar integration for the higher moments can be done in a similar manner. Here, we recall some useful single-qubit Haar integration formulae

(see, for example, Eq. (2.26) in Ref. [7])

$$\int dV V^{\otimes 1} |0\rangle\langle 0|^{\otimes 1} V^{\dagger \otimes 1} = \frac{1}{2} \mathbb{1} \quad (152)$$

$$\int dV V^{\otimes 2} |0\rangle\langle 0|^{\otimes 2} V^{\dagger \otimes 2} = \frac{1}{6} (\mathbb{1} \otimes \mathbb{1} + S_{12}) \quad (153)$$

$$\int dV V^{\otimes 3} |0\rangle\langle 0|^{\otimes 3} V^{\dagger \otimes 3} = \frac{1}{24} (\mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} + S_{12} + S_{13} + S_{23} + S_{23}S_{12} + S_{23}S_{13}) , \quad (154)$$

$$\begin{aligned} \int dV V^{\otimes 4} |0\rangle\langle 0|^{\otimes 4} V^{\dagger \otimes 4} = & \frac{1}{120} (\mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} + S_{12} + S_{13} + S_{14} + S_{23} + S_{24} + S_{34} + S_{34}S_{12} + S_{24}S_{13} + S_{23}S_{14} \\ & + S_{23}S_{12} + S_{24}S_{12} + S_{23}S_{13} + S_{34}S_{13} + S_{24}S_{14} + S_{34}S_{23} + S_{34}S_{24} + S_{34}S_{41} \\ & + S_{34}S_{23}S_{12} + S_{34}S_{24}S_{12} + S_{24}S_{23}S_{13} + S_{24}S_{34}S_{13} + S_{23}S_{34}S_{14} + S_{23}S_{24}S_{14}) \end{aligned} \quad (155)$$

where S_{lk} is the swap operator between systems l and k .

In addition, we will use the following lemma for the variance of an arbitrary operator O in the Pauli basis over random product states.

Lemma 1. *Consider an arbitrary observable O decomposed into the Pauli basis*

$$O = \sum_{\sigma \in \mathbf{p}_n} \lambda_\sigma \sigma, \quad (156)$$

where the weights λ_σ are real constants and $\mathbf{p}_n = \{\mathbb{1}, X, Y, Z\}^{\otimes n}$ is the Pauli ensemble on n qubits. The variance of O over single qubit Haar random product states is given by

$$\text{Var}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}[O] = \sum_{\sigma \in \mathbf{p}_n \setminus \{\mathbb{1}^{\otimes n}\}} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}} , \quad (157)$$

where $s(\sigma)$ is the subset of qubits on which σ acts non trivially and $|s(\sigma)|$ is a cardinality of $s(\sigma)$.

Proof. We consider the arbitrary observable O which can be deomposed into the Pauli basis as

$$O = \sum_{\sigma \in \mathbf{p}_n} \lambda_\sigma \sigma, \quad (158)$$

where the weights λ_σ are real constants and $\mathbf{p}_n = \{\mathbb{1}, X, Y, Z\}^{\otimes n}$ is the Pauli ensemble on n qubits. We denote $s(\sigma)$ as a support of σ which is a subset of qubits that σ acts non-trivially on and $|s(\sigma)|$ as a cardinality of $s(\sigma)$ ¹.

The variance of O over single qubit Haar random product states $|\psi\rangle \sim \text{Haar}_1^{\otimes n}$ is of the form

$$\text{Var}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}[O] = \mathbb{E}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} [|\langle \psi | O | \psi \rangle|^2] - \left(\mathbb{E}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} [|\langle \psi | O | \psi \rangle|] \right)^2 \quad (159)$$

$$= \langle O^2 \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} - \langle O \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}^2 . \quad (160)$$

First, we consider the average of O over single qubit Haar random product states.

$$\langle O \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} = \sum_{\sigma \in \mathbf{p}_n} \lambda_\sigma \langle \sigma \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} \quad (161)$$

$$= \sum_{\sigma \in \mathbf{p}_n} \lambda_\sigma \prod_{i=1}^n \langle \sigma_i \rangle_{|\psi_i\rangle \sim \text{Haar}_1} \quad (162)$$

$$= \sum_{\sigma \in \mathbf{p}_n} \lambda_\sigma \prod_{i=1}^n \delta(\sigma_i = \mathbb{1}) \quad (163)$$

$$= \lambda_{\mathbb{1}^{\otimes n}} , \quad (164)$$

¹ As an example, for $\sigma = X \otimes \mathbb{1} \otimes \mathbb{1} \otimes Z \otimes Y$, we have $s(\sigma) = \{1, 4, 5\}$ with $|s(\sigma)| = 3$.

where, in the third equality, we use the Haar integration formula in Eq. (152) together with the fact that all single-qubit Pauli matrices are traceless, and we denote $\delta(\sigma_i = \mathbb{1}) = 1$ if $\sigma_i = \mathbb{1}$ (otherwise, $\delta(\sigma_i = \mathbb{1}) = 0$).

Now, we consider the second-moment of O over a random product state. To evaluate this we follow the proof of Lemma B.3 in Appendix B in Ref. [8] to integrate over the random product states but replace the unitary $\tilde{U}^\dagger W \hat{U}$ in Ref. [8] with an observable O . This directly leads to

$$\langle O^2 \rangle = \frac{1}{6^n} \sum_{A \subseteq \mathcal{N}} \text{Tr}[O_A^2], \quad (165)$$

where $O_A = \text{Tr}_{\bar{A}}[O]$ is the partial trace of O over all qubits except those in $A \subseteq \mathcal{N} = \{1, 2, \dots, n\}$. We recall that A is also defined in Eq. (110).

Consider a given subset of qubits A . We first notice that, for a given $\sigma = \bigotimes_{i=1}^n \sigma_i$, we have the partial trace of the Pauli string over A as

$$\sigma_A = \text{Tr}_{\bar{A}} \left[\bigotimes_{i=1}^n \sigma_i \right] \quad (166)$$

$$= \left[\prod_{i \notin A} \text{Tr}[\sigma_i] \right] \cdot \left[\bigotimes_{i \in A} \sigma_i \right] \quad (167)$$

$$= 2^{n-|A|} \delta(s(\sigma) \subseteq A) \bigotimes_{i \in A} \sigma_i, \quad (168)$$

where we denote $\delta(s(\sigma) \subseteq A) = 1$ if $s(\sigma) \subseteq A$ and $\delta(s(\sigma) \subseteq A) = 0$ if $s(\sigma) \not\subseteq A$, which is a direct consequence of the Pauli matrices being traceless i.e., $\text{Tr}[\sigma_i] \neq 0$ only if $\sigma_i = \mathbb{1}$. Importantly, $\sigma_A \neq 0$ only if the part that σ acts non-trivially is a subset of A .

Now, we consider

$$\text{Tr}[O_A^2] = \text{Tr}_A [\text{Tr}_{\bar{A}}[O] \cdot \text{Tr}_{\bar{A}}[O]] \quad (169)$$

$$= \text{Tr} \left[\left(\sum_{\sigma \in \mathbf{p}_n} \lambda_\sigma 2^{n-|A|} \delta(s(\sigma) \subseteq A) \bigotimes_{i \in A} \sigma_i \right) \left(\sum_{\sigma' \in \mathbf{p}_n} \lambda_{\sigma'} 2^{n-|A|} \delta(s(\sigma') \subseteq A) \bigotimes_{i \in A} \sigma'_i \right) \right] \quad (170)$$

$$= \sum_{\sigma, \sigma' \in \mathbf{p}_n} \lambda_\sigma \lambda_{\sigma'} 2^{2(n-|A|)} \delta(s(\sigma) \subseteq A) \delta(s(\sigma') \subseteq A) \left(\prod_{i \in A} \text{Tr}[\sigma_i \sigma'_i] \right) \quad (171)$$

$$= \sum_{\sigma, \sigma' \in \mathbf{p}_n} \lambda_\sigma \lambda_{\sigma'} 2^{2(n-|A|)} \delta(s(\sigma) \subseteq A) \delta(s(\sigma') \subseteq A) \left(2^{|A|} \prod_{i \in A} \delta(\sigma_i = \sigma'_i) \right) \quad (172)$$

$$= \sum_{\sigma \in \mathbf{p}_n} \lambda_\sigma^2 2^{2n-|A|} \delta(s(\sigma) \subseteq A), \quad (173)$$

where the third equality is due to $\text{Tr}[(\bigotimes_{i \in A} \sigma_i) (\bigotimes_{i \in A} \sigma'_i)] = \prod_{i \in A} \text{Tr}[\sigma_i \sigma'_i]$, and the fourth equality is due to $\text{Tr}[\sigma_i \sigma'_i] = 2\delta(\sigma_i = \sigma'_i)$ with $\delta(\sigma_i = \sigma'_i) = 1$ if $\sigma_i = \sigma'_i$ and $\delta(\sigma_i = \sigma'_i) = 0$, otherwise. In the last equality, we notice that the condition that σ acts non-trivially only on A (i.e., $\delta(s(\sigma) \subseteq A)$) together with the reduced Pauli strings on A are the same for σ and σ' (i.e., $\prod_{i \in A} \delta(\sigma_i = \sigma'_i)$) implies that $\sigma = \sigma'$ for the term to be non-zero, reducing the double sum to the single sum.

We are ready to continue with

$$\langle O^2 \rangle = \frac{1}{6^n} \sum_{A \subseteq \mathcal{N}} \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 2^{2n-|A|} \delta(s(\sigma) \subseteq A) \quad (174)$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 \sum_{A \subseteq \mathcal{N}} 2^{-|A|} \delta(s(\sigma) \subseteq A) \quad (175)$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 \sum_{A' \subseteq \mathcal{N} \setminus s(\sigma)} 2^{-|s(\sigma)|-|A'|} \quad (176)$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 2^{-|s(\sigma)|} \sum_{|A'|=0}^{n-|s(\sigma)|} \binom{n-|s(\sigma)|}{|A'|} 2^{-|A'|} \quad (177)$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 2^{-|s(\sigma)|} \left(\frac{3}{2}\right)^{n-|s(\sigma)|} \quad (178)$$

$$= \sum_{\sigma \in \mathfrak{p}_n} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}} , \quad (179)$$

where the third equality is due to the fact that the terms do not vanish only when $s(\sigma) \in A$ and therefore we only have to sum over A that contain $s(\sigma)$. The latter is equivalent to summing A' where $A = A' \cup s(\sigma)$ over $\mathcal{N} \setminus s(\sigma)$. In fourth equality, we replace the sum over A' by a sum over $|A'|$ and counted the number of ensembles of size $|A'|$ in $\mathcal{N} \setminus s(\sigma)$ (which is of size $n - |s(\sigma)|$). In the fifth equality, we recognised a binomial sum.

Lastly, we have the variance of the form

$$\text{Var}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}[O] = \sum_{\sigma \in \mathfrak{p}_n} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}} - (\lambda_{\mathbb{1}^{\otimes n}})^2 \quad (180)$$

$$= \sum_{\sigma \in \mathfrak{p}_n \setminus \mathbb{1}^{\otimes n}} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}} , \quad (181)$$

where the sum in the last line excludes the identity term. This completes the proof of the lemma. $\square$

2. Generic form of the MMD variance for a tensor product ansatz

We now give a generic expression of the variance of the MMD loss for an arbitrary bandwidth, which is stated in the following proposition

Supplemental Proposition 2. Consider the MMD loss function $\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})$ as defined in Eq. (13) in the main text, which uses the classical Gaussian kernel as defined in Eq. (101) with the bandwidth σ , and a quantum generative model that is comprised of a tensor-product ansatz $U = \bigotimes_i^n U_i(\theta_i)$ with $\{U_i(\theta_i)\}_{\theta_i}$ a single-qubit Haar random ensemble for all i . Given a training dataset $\tilde{P}$, we have that the variance of the MMD loss over parameters $\boldsymbol{\theta}$ is

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] = B_\sigma + 4C_\sigma(\tilde{P}) , \quad (182)$$

with

$$B_\sigma = \left[\frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right]^n - \left[\frac{4 + 4e^{-1/2\sigma} + e^{-1/\sigma}}{9} \right]^n , \quad (183)$$

and

$$C_\sigma(\tilde{P}) = \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (1 - p_\sigma)^{2(n-|A|)} \left(\frac{p_\sigma^2}{3}\right)^{|A|} z_A^2(\tilde{P}) , \quad (184)$$

where $p_\sigma = (1 - e^{-\frac{1}{2\sigma}})/2$, $\mathcal{N} = \{1, 2, \dots, n\}$, $z_A(\tilde{P}) = \text{Tr}[(\bigotimes_{i \in A} Z_i)\rho_{\tilde{P}}] = \sum_{\mathbf{y}} \tilde{p}(\mathbf{y})(-1)^{\sum_{i \in A} y_i}$ with $\rho_{\tilde{P}}$ being the quantum state corresponding to the training data such that $\tilde{p}(\mathbf{x}) = \text{Tr}[\rho_{\tilde{P}}|\mathbf{x}\rangle\langle\mathbf{x}|]$. The sum in Eq. (184) is over all possible subsets of $\mathcal{N}$ excluding the empty set $\{\}$.

We remark that B_σ and $C_\sigma(\tilde{P})$ are the variances of the first term $\mathcal{K}_{q,q}(\boldsymbol{\theta})$ and the second term $\mathcal{K}_{p,q}(\boldsymbol{\theta})$ in the MMD, respectively, while we found the covariance term to vanish. The dependence on the training data is encoded in $z_A(\tilde{P})$ which ranges between -1 and 1 . Lastly, the exact formula of the MMD variance has been found to be consistent with the numerical simulation up to $n = 1000$ in Fig. 6 in the main text.

Proof. There are three main steps in our proof. (i) computing the variance of the first term $\mathcal{K}_{q,q}(\boldsymbol{\theta})$, (ii) computing the variance of the second term $\mathcal{K}_{p,q}(\boldsymbol{\theta})$ and, lastly, (iii) showing that the covariance between the two terms is zero.

(i) Computing the variance of $\mathcal{K}_{q,q}(\boldsymbol{\theta})$:

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}^2(\boldsymbol{\theta})] - (\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})])^2 \quad (185)$$

$$= \mathbb{E}_{\boldsymbol{\theta}} \left[\left(\sum_{\mathbf{x}, \mathbf{y}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) \right)^2 \right] - \left(\mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{\mathbf{x}, \mathbf{y}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) \right] \right)^2 \quad (186)$$

$$= \sum_{\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'} \mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) q_{\boldsymbol{\theta}}(\mathbf{x}') q_{\boldsymbol{\theta}}(\mathbf{y}')] K(\mathbf{x}, \mathbf{y}) K(\mathbf{x}', \mathbf{y}') - \left(\sum_{\mathbf{x}, \mathbf{y}} \mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y})] K(\mathbf{x}, \mathbf{y}) \right)^2. \quad (187)$$

We now can express each individual model probability as in Eq. (147) and then average over the parameters $\boldsymbol{\theta}$. This requires us to perform Haar integration for the first and second terms in Eq. (187), respectively.

First, consider

$$\sum_{\mathbf{x}, \mathbf{y}} \mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y})] K(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \int dU_i(\boldsymbol{\theta}_i) \text{Tr}[(U_i(\boldsymbol{\theta}_i))^{\otimes 2} |0_i\rangle\langle 0_i|^{\otimes 2} (U_i^\dagger(\boldsymbol{\theta}_i))^{\otimes 2} (|x_i\rangle\langle x_i| \otimes |y_i\rangle\langle y_i|)] K(\mathbf{x}, \mathbf{y}) \quad (188)$$

$$= \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \text{Tr} \left[\left(\int dU_i(\boldsymbol{\theta}_i) (U_i(\boldsymbol{\theta}_i))^{\otimes 2} |0_i\rangle\langle 0_i|^{\otimes 2} (U_i^\dagger(\boldsymbol{\theta}_i))^{\otimes 2} \right) (|x_i\rangle\langle x_i| \otimes |y_i\rangle\langle y_i|) \right] K(\mathbf{x}, \mathbf{y}) \quad (189)$$

$$= \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \text{Tr} \left[\left(\frac{\mathbb{1} \otimes \mathbb{1} + S_{12}}{6} \right) (|x_i\rangle\langle x_i| \otimes |y_i\rangle\langle y_i|) \right] e^{-\frac{(x_i - y_i)^2}{2\sigma}} \quad (190)$$

$$= \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \left(\frac{1 + \delta_{x_i, y_i}}{6} \right) e^{-\frac{(x_i - y_i)^2}{2\sigma}} \quad (191)$$

$$= \sum_{\mathbf{x}} \prod_{i=1}^n \left[\left(\frac{1 + \delta_{x_i, 0}}{6} \right) e^{-\frac{(x_i)^2}{2\sigma}} + \left(\frac{1 + \delta_{x_i, 1}}{6} \right) e^{-\frac{(x_i - 1)^2}{2\sigma}} \right] \quad (192)$$

$$= \left(\frac{2 + e^{-1/2\sigma}}{3} \right)^n, \quad (193)$$

where, in the third equality, we use Eq. (153), and, in the fifth equality as well as in the last equality, we use the identity $\sum_{\mathbf{x}} \prod_{i=1}^n h_i(x_i) = \prod_{i=1}^n (h_i(0) + h_i(1))$.

Similarly, the first term in Eq. (187) can be computed via Haar integration using Eq. (155) and repeatedly applying the identity $\sum_{\mathbf{x}} \prod_{i=1}^n h_i(x_i) = \prod_{i=1}^n (h_i(0) + h_i(1))$, leading to

$$\sum_{\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'} \mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) q_{\boldsymbol{\theta}}(\mathbf{x}') q_{\boldsymbol{\theta}}(\mathbf{y}')] K(\mathbf{x}, \mathbf{y}) K(\mathbf{x}', \mathbf{y}') = \left(\frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right)^n. \quad (194)$$

Altogether, we have the variance of the first MMD term as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})] = \left(\frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right)^n - \left(\frac{4 + 4e^{-1/2\sigma} + e^{-1/\sigma}}{9} \right)^n. \quad (195)$$

(ii) Computing the variance of $\mathcal{K}_{p,q}(\boldsymbol{\theta})$: There are two alternative ways of doing this, leading to two equivalent expressions of the variance of $\mathcal{K}_{p,q}(\boldsymbol{\theta})$. First is the same approach used in (i). Alternatively, we can interpret the middle term as an expectation value of an observable $O_{\tilde{p}}^{(\sigma)}(\tilde{P}) = \sum_{\mathbf{x}} \lambda_{\mathbf{x}}(\tilde{P}) |\mathbf{x}\rangle\langle\mathbf{x}|$ with $\lambda_{\mathbf{x}}(\tilde{P}) = \sum_{\mathbf{y}} p(\mathbf{y}) K(\mathbf{x}, \mathbf{y})$ and then use Lemma 1. That is, $\mathcal{K}_{p,q}(\boldsymbol{\theta}) = \text{Tr}[U(\boldsymbol{\theta})|\mathbf{0}\rangle\langle\mathbf{0}|U^\dagger(\boldsymbol{\theta})O_{\tilde{p}}^{(\sigma)}(\tilde{P})]$. Transforming $O_{\tilde{p}}^{(\sigma)}$ into the Pauli basis with $|\mathbf{x}\rangle\langle\mathbf{x}| = \bigotimes_{i=1}^n |x_i\rangle\langle x_i| = \bigotimes_{i=1}^n \frac{1}{2}(\mathbb{1}_i + (-1)^{x_i} Z_i)$ leads to

$$O_{\tilde{p}}^{(\sigma)} = \sum_{A \subseteq \mathcal{N}} (1 - p_\sigma)^{n-|A|} p_\sigma^{|A|} z_A(\tilde{P}) \bigotimes_{i \in A} Z_i, \quad (196)$$

where $\mathcal{N} = \{1, 2, \dots, n\}$, $p_\sigma = (1 - e^{-\frac{1}{2\sigma}})/2$ and we denote

$$z_A(\tilde{P}) = \text{Tr}\left[\left(\bigotimes_{i \in A} Z_i\right) \rho_{\tilde{p}}\right] \quad (197)$$

$$= \sum_{\mathbf{y}} \tilde{p}(\mathbf{y}) (-1)^{\sum_{i \in A} y_i}, \quad (198)$$

where $\rho_{\tilde{p}}$ is the quantum state associated with the training data with $\tilde{p}(\mathbf{x}) = \text{Tr}[\rho_{\tilde{p}}|\mathbf{x}\rangle\langle\mathbf{x}|]$ ².

By using Lemma 1, the variance of the middle term with $O_{\tilde{p}}^{(\sigma)}$ expressed in the Pauli basis is of the form

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{p,q}(\boldsymbol{\theta})] = \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (1 - p_\sigma)^{2(n-|A|)} \left(\frac{p_\sigma^2}{3}\right)^{|A|} z_A^2(\tilde{P}). \quad (199)$$

Notice that the sum now excludes the empty set $\{\}$. We can see that $z_A(\tilde{P})$ encodes information about the target distribution.

(iii) Computing the covariance between $\mathcal{K}_{p,p}(\boldsymbol{\theta})$ and $\mathcal{K}_{p,q}(\boldsymbol{\theta})$: By direct computation as in (i), we have

$$\text{Cov}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta}), \mathcal{K}_{p,q}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})\mathcal{K}_{p,q}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})]\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{K}_{p,q}(\boldsymbol{\theta})] \quad (200)$$

$$= \sum_{\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'} p(\mathbf{y}') \left(\mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})q_{\boldsymbol{\theta}}(\mathbf{y}')] - \mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})]\mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x}')] \right) K(\mathbf{x}, \mathbf{y})K(\mathbf{x}', \mathbf{y}') \quad (201)$$

$$= 0, \quad (202)$$

where the last equality follows from

$$\sum_{\mathbf{x}, \mathbf{y}} \left(\mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})q_{\boldsymbol{\theta}}(\mathbf{y}')] - \mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})]\mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x}')] \right) K(\mathbf{x}, \mathbf{y}) = 0 \quad (203)$$

which holds for any $\mathbf{x}'$ and $\mathbf{y}'$ from Eq. (148), Eq. (153) and Eq. (154).

By substituting Eq. (195), Eq. (199) and Eq. (202) back into the MMD variance expression in Eq. (142), the proof is completed. $\square$

² Equivalently, one can get to this reduced MMD observable by tracing out half of the qubits that $\rho_{\tilde{p}}$ acts on in $O_{\text{MMD}}^{(\sigma)}$ in Eq. (108). That is, $O_{\tilde{p}}^{(\sigma)} = \text{Tr}_1[(\mathbb{1} \otimes \rho_{\tilde{p}})O_{\text{MMD}}^{(\sigma)}]$.

3. Variance scaling and trainability of MMD

We now analyze how the scaling of the variance depends on the bandwidth σ . To demonstrate the presence of loss concentration, it is sufficient to show that the variance of the whole MMD loss has an exponentially small upper bound. We show that this happens when the bandwidth is constant and independent of the number of qubits, i.e., $\sigma \in \mathcal{O}(1)$. On the other hand, to establish trainability it is crucial to accurately measure all individual terms in the MMD loss. More precisely, we require both first and second MMD terms to have at least a polynomially large variance. We argue that this can be achieved by using bandwidth that scales as $\sigma \in \Theta(n)$.

A formal version of Theorem 2 in the main text is stated below.

Theorem 2 (Product ansatz trainability of MMD, formal). *Consider the MMD loss function $\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})$ as defined in Eq. (99), which uses the classical Gaussian kernel as defined in Eq. (101) with the bandwidth $\sigma > 0$, and a quantum circuit generative model that is comprised of a tensor-product ansatz $U = \bigotimes_i^n U_i(\theta_i)$ with $\{U_i(\theta_i)\}_{\theta_i}$ being single-qubit (Haar) random unitaries. Given a training dataset $\tilde{P}$, the asymptotic scaling of the variance of the MMD loss depends on the value of σ .*

For $\sigma \in \mathcal{O}(1)$, we have

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] \in \mathcal{O}(1/b^n), \quad (204)$$

with some $b > 1$.

On the other hand, according to Supplemental Proposition 2 and for $\sigma \in \Theta(n)$, we have

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] = B_{\sigma} + 4C_{\sigma}(\tilde{P}), \quad (205)$$

where the variance of the first term B_{σ} is lower-bounded as

$$B_{\sigma} \in \Omega(1/n), \quad (206)$$

as well as, the variance of the second term $C_{\sigma}(\tilde{P})$ is lower-bounded as

$$C_{\sigma}(\tilde{P}) \in \Omega(1/\text{poly}(n)), \quad (207)$$

provided that

$$\sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}) \in \Omega(1/\text{poly}(n)), \quad (208)$$

with $k \in \mathcal{O}(1)$, $\mathcal{N} = \{1, \dots, n\}$ and $z_A(\tilde{P}) = \sum_{\mathbf{y}} \tilde{p}(\mathbf{y}) (-1)^{\sum_{i \in A} y_i}$ which encodes the information about the training data.

Proof. We consider the scaling of the MMD variance in Eq. (182) for two scenarios of the bandwidth values.

(i) For $\sigma \in \mathcal{O}(1)$: We show that both B_{σ} and $C_{\sigma}(\tilde{P})$ in the MMD variance are exponentially small. First, from Eq. (183) we know that

$$B_{\sigma} \leq \left[\frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right]^n. \quad (209)$$

If $\sigma \in \mathcal{O}(1)$, then $\tilde{B}_{\sigma}$ is exponentially decreasing as $e^{-1/2\sigma}, e^{-1/\sigma} \in \mathcal{O}(1)$.

For the second MMD term, Eq. (184), we upper bound as

$$C_\sigma(\tilde{P}) \leq \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (1 - p_\sigma)^{2(n-|A|)} \left(\frac{p_\sigma^2}{3} \right)^{|A|} \quad (210)$$

$$= \sum_{|A|=1}^n \binom{n}{|A|} (1 - p_\sigma)^{2(n-|A|)} \left(\frac{p_\sigma^2}{3} \right)^{|A|} \quad (211)$$

$$\leq \sum_{|A|=0}^n \binom{n}{|A|} (1 - p_\sigma)^{2(n-|A|)} \left(\frac{p_\sigma^2}{3} \right)^{|A|} \quad (212)$$

$$= \left((1 - p_\sigma)^2 + \frac{p_\sigma^2}{3} \right)^n \quad (213)$$

$$= \left[\frac{1 + e^{-1/2\sigma} + e^{-1/\sigma}}{3} \right]^n, \quad (214)$$

where the first inequality is due to $z_A^2(\tilde{P}) \leq 1$, the first equality leverages that the expression only depends on the support of A rather than A itself, the final inequality is due to including the empty set in the sum (i.e., $|A| = 0$), and in the second equality we use the binomial sum formula. In the final line we use the definition of p_σ from Eq. (109). Similarly to the first term, if $\sigma \in \mathcal{O}(1)$, the upper bound decays exponentially with n .

Therefore, when $\sigma \in \mathcal{O}(1)$, the variance of the MMD loss scales as

$$\text{Var}_\theta[\mathcal{L}_{\text{MMD}}(\theta)] \in \mathcal{O}(1/b^n), \quad (215)$$

for some $b > 0$.

(ii) For $\sigma \in \Theta(n)$: Consider the variance of the first MMD term

$$B_\sigma = \left[\frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right]^n - \left[\frac{4 + 4e^{-1/2\sigma} + e^{-1/\sigma}}{9} \right]^n \quad (216)$$

$$\geq \left[\frac{7 + 6\left(1 - \frac{1}{2\sigma} + \frac{1}{8\sigma^2} - \frac{1}{48\sigma^3}\right) + 2\left(1 - \frac{1}{\sigma} + \frac{1}{2\sigma^2} - \frac{1}{6\sigma^3}\right)}{15} \right]^n - \left[\frac{4 + 4\left(1 - \frac{1}{2\sigma} + \frac{1}{8\sigma^2}\right) + \left(1 - \frac{1}{\sigma} + \frac{1}{2\sigma^2}\right)}{9} \right]^n \quad (217)$$

$$= \left[1 - \frac{1}{3\sigma} + \frac{7}{60\sigma^2} - \frac{11}{360\sigma^3} \right]^n - \left[1 - \frac{1}{3\sigma} + \frac{1}{9\sigma^2} \right]^n \quad (218)$$

$$= \left(1 - \frac{1}{3\sigma} \right)^n \left[\left[1 + \frac{\frac{7}{60\sigma^2} - \frac{11}{360\sigma^3}}{1 - \frac{1}{3\sigma}} \right]^n - \left[1 + \frac{\frac{1}{9\sigma^2}}{1 - \frac{1}{3\sigma}} \right]^n \right] \quad (219)$$

$$\geq \left(1 - \frac{1}{3\sigma} \right)^{n-1} \left(\frac{n}{180\sigma^2} \right) \left(1 - \frac{11}{2\sigma} \right) \quad (220)$$

$$= \left(\frac{n}{180\sigma^2} \right) \left(1 - \frac{11}{2\sigma} \right) \left[\left(1 - \frac{1}{3\sigma} \right)^{3\sigma} \right]^{(n-1)/3\sigma}, \quad (221)$$

where in the first inequality we use $1 - x + x^2/2 - x^3/6 \leq e^{-x} \leq 1 - x + x^2/2$, the second inequality is due to $(1+a)^n - (1+b)^n \geq n(a-b)$ for positive a, b and $a > b$. This is satisfied when $\sigma > 5.5$, which is the case for sufficiently large n . We note that $(n-1)/3\sigma \in \mathcal{O}(1)$. To proceed further, we consider the following lemma.

Lemma 2. The lower bound of $f(x) = (1 - 1/x)^x$ with $1 < |x|$ is given by

$$f(x) \geq \frac{1}{e} \left(1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)x^j} \right). \quad (222)$$

Proof. We consider

$$f(x) = \exp(x \ln(1 - 1/x)) \quad (223)$$

$$= \exp\left(x \left(\sum_{j=1}^{\infty} -\frac{1}{j x^j}\right)\right) \quad (224)$$

$$= \frac{1}{e} \cdot \exp\left(\sum_{j=1}^{\infty} -\frac{1}{(j+1)x^j}\right) \quad (225)$$

$$\geq \frac{1}{e} \left(1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)x^j}\right), \quad (226)$$

where the second equality is due to the Taylor expansion of $\ln(1 - 1/x)$ which converges for $1 < |x|$, the inequality is by using $e^{-y} \geq 1 - y$. $\square$

By using Lemma 2, we have the following lower bound

$$B_{\sigma} \geq \left(\frac{n}{180\sigma^2}\right) \left(1 - \frac{11}{2\sigma}\right) \left[\frac{1}{e} \left(1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)(3\sigma)^j}\right)\right]^{(n-1)/3\sigma}, \quad (227)$$

which implies that $B_{\sigma} \in \Omega(1/n)$ for $\sigma \in \Omega(n)$

Similarly, for the second term, we have

$$C_{\sigma}(\tilde{P}) = \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}}} (1 - p_{\sigma})^{2(n-|A|)} \left(\frac{p_{\sigma}^2}{3}\right)^{|A|} z_A^2(\tilde{P}) \quad (228)$$

$$= \left[\frac{1 + 2e^{-1/2\sigma} + e^{-1/\sigma}}{4}\right]^n \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}}} \left(\frac{\tanh^2(1/4\sigma)}{3}\right)^{|A|} z_A^2(\tilde{P}) \quad (229)$$

$$\geq \left[\frac{1 + 2\left(1 - \frac{1}{2\sigma}\right) + \left(1 - \frac{1}{\sigma}\right)}{4}\right]^n \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}}} \left(\frac{\frac{1}{16\sigma^2} \left(1 - \frac{1}{24\sigma^2}\right)}{3}\right)^{|A|} z_A^2(\tilde{P}) \quad (230)$$

$$\geq \left[\left(1 - \frac{1}{2\sigma}\right)^{2\sigma}\right]^{\frac{n}{2\sigma}} \left(\frac{1}{48\sigma^2} \left(1 - \frac{1}{24\sigma^2}\right)\right)^k \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}) \quad (231)$$

$$\geq \left[\frac{1}{e} \left(1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)(2\sigma)^j}\right)\right]^{\frac{n}{2\sigma}} \left(\frac{1}{48\sigma^2} \left(1 - \frac{1}{24\sigma^2}\right)\right)^k \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}), \quad (232)$$

where $\tanh(1/4\sigma) = p_{\sigma}/(1 - p_{\sigma})$, the first inequality is due to $e^{-x} \geq 1 - x$ and $\tanh(x) \geq x - x^3/3$ (for positive x), and in the second inequality we truncate the sum at $|A| = k$ and taking $|A| = k$ for all terms within the truncated sum. Finally, in the last inequality, we note that $\frac{n}{2\sigma} \in \mathcal{O}(1)$ and use Lemma 2.

Altogether, taking $k \in \mathcal{O}(1)$ and assuming

$$\sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}) \in \Omega(1/\text{poly}(n)), \quad (233)$$

the MMD variance is lower bounded as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}(\boldsymbol{\theta})] \in \Omega(1/n), \quad (234)$$

with the desired scaling of B_{σ} and $C_{\sigma}(\tilde{P})$. This completes the proof of the theorem. $\square$

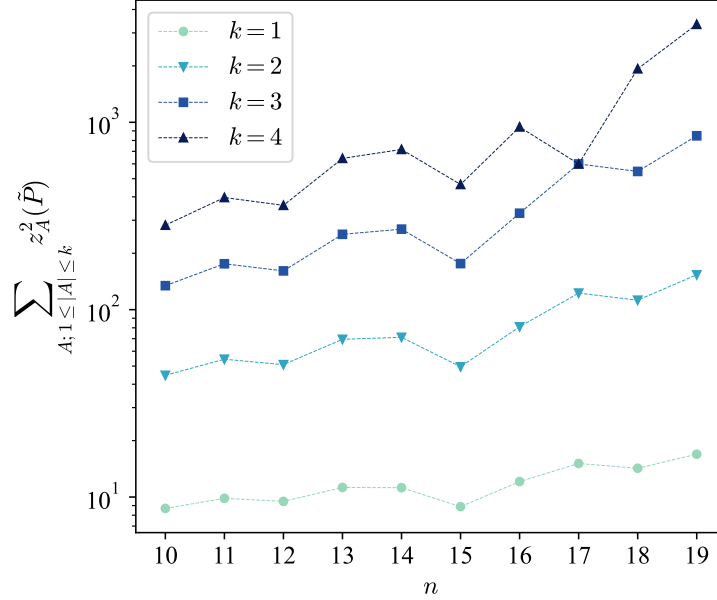


Figure 1. We plot $\sum_{A \subseteq \mathcal{N}} z_A^2(\tilde{P})$ such that $|A| \leq k$ and $A \neq \{\}$ as a function of n and shows different k values. The datasets considered here corresponds to the real high energy physics dataset used in Sec. “Training on a HEP dataset” in the main text. Values of $n = 10, 11, \dots, 19$ and $k = 1, 2, 3, 4$ are presented. The quantity does not vanish with the increasing number of qubits, satisfying the data-dependence assumption made in Theorem 2.

We showed that the variance of the MMD cross terms $C_{\sigma}(\tilde{P})$ decays at most polynomially in n provided that the sum of $z_A^2(\tilde{P})$ terms for $|A| \in \mathcal{O}(1)$ and $A \neq \{\}$ is at least polynomially small in n , i.e., not exponentially small. Here, we comment on this assumption.

First, we recall the definition of $z_A(\tilde{P})$ from Eq. (197)

$$z_A(\tilde{P}) = \text{Tr} \left[\left(\bigotimes_{i \in A} Z_i \right) \rho_{\tilde{P}} \right] \quad (235)$$

$$= \sum_{\mathbf{y}} \tilde{p}(\mathbf{y}) (-1)^{\sum_{i \in A} y_i}, \quad (236)$$

with $\rho_{\tilde{P}}$ being the quantum state associated with the training data distribution $\tilde{p}(\mathbf{x}) = \text{Tr}[\rho_{\tilde{P}}|\mathbf{x}\rangle\langle\mathbf{x}|]$. Importantly, $z_A(\tilde{P})$ encodes the correlation of training data on the subset A of the bitstring (with A defined in Eq. (110)), which can be interpreted as an average parity over A . In other words, its purpose is for the model to learn the same expectation on the operator Z_A as the dataset.

The magnitude of $\sum_A z_A^2(\tilde{P})$ depends on the provided training dataset, and if the variance of the cross term vanishes due to the data-dependence, this is because it is required for a faithfulness of the loss function. We note, however, that we do not expect this to occur in practice because partial datasets are likely exhibit significant correlations, and also

due to the assumption of polynomial dataset sizes, i.e., $\tilde{p}(\mathbf{x}) \in \Omega(1/\text{poly}(n))$ for $\mathbf{x} \in \tilde{P}$. To emphasize this point, we analyzed the dataset from HEP colliders experiments used throughout Sec. “Training on a HEP dataset” in the main text. Fig. 1 numerically shows the sum of the first $z_A^2(\tilde{P})$ terms for $1 \leq |A| \leq k$ as a function of n and for $k = 1, 2, 3$ and 4. We see that these terms for the low-body interactions do not disappear as the number of qubits increases and, in fact, they increase with n instead. This provides a practical example that the data-dependence assumption used in Theorem 2 is satisfied in practice.

C. MMD variance for a generic Pauli rotation ansätze

In this sub-section, we present a formal version of Theorem 3 and provide details of the proof. First, we recall that the MMD observable in Eq. (107) is of the form

$$O_{\text{MMD}}^{(\sigma)} = \sum_{A \in \mathcal{N}} (1 - p_\sigma)^{n-|A|} p_\sigma^{|A|} \bigotimes_{i \in A} (Z_i \otimes Z_{n+i}) \quad (237)$$

$$= \sum_{A \in \mathcal{N}} c_\sigma(A) Z_A \otimes Z_A \quad (238)$$

where we denote A as the set of all possible subsets of the indices $1, \dots, n$ as defined in Eq. (110), and in the second equality we introduce some shorthand notations $c_\sigma(A) = (1 - p_\sigma)^{n-|A|} p_\sigma^{|A|}$ and $Z_A = \bigotimes_{i \in A} Z_i$. Also, each term in the MMD loss function can be expressed as some expectation of the form

$$\mathcal{M}(\rho, \rho') = \text{Tr} \left[O_{\text{MMD}}^{(\sigma)} (\rho \otimes \rho') \right], \quad (239)$$

and the MMD loss function is of the form

$$\mathcal{L}_{\text{MMD}}(\boldsymbol{\theta}) = \mathcal{M}(\rho_\theta, \rho_\theta) - 2\mathcal{M}(\rho_\theta, \rho_{\tilde{p}}) + \mathcal{M}(\rho_{\tilde{p}}, \rho_{\tilde{p}}), \quad (240)$$

where $\rho_\theta = U(\boldsymbol{\theta})\rho_0 U^\dagger(\boldsymbol{\theta})$ and $\rho_{\tilde{p}}$ is a quantum state such that $\tilde{p}(\mathbf{x}) = \text{Tr}[\rho_{\tilde{p}}|\mathbf{x}\rangle\langle\mathbf{x}|]$. For those who want more refresher on the MMD observable, we refer to Supplementary Note IV A.

We consider a generic Pauli rotation ansätze of the form

$$U_{\text{PQC}}(\boldsymbol{\theta}) = U(\boldsymbol{\alpha}) U_{\text{tensor}}(\boldsymbol{\beta}), \quad (241)$$

where $U(\boldsymbol{\alpha}) = \prod_{k=1}^M e^{-i\alpha_k G_k/2} V_k$ such that the generators $\{G_k\}_{k=1}^M$ are some n -qubit Pauli strings $G_k \in \mathfrak{p}_n := \{\mathbb{1}, X, Y, Z\}^n$ and $\{V_k\}_{k=1}^M$ is a set of non-parametrized Clifford gates, and $U_{\text{tensor}}(\boldsymbol{\beta}) = \bigotimes_{i=1}^n U_i(\beta_i)$ is a layer of Haar random single qubit rotations $U_i(\beta_i)$. In addition, we assume all parameters are uncorrelated.

Before showing the main theoretical results, we remarks that any arbitrary Pauli-string $B \in \mathfrak{p}_n := \{\mathbb{1}, X, Y, Z\}^{\otimes n}$ remains a single Pauli-string when being back-propagated with $U(\boldsymbol{\alpha}_c)$ where $\boldsymbol{\alpha} = \boldsymbol{\alpha}_c \in \{0, \pi/2\}^M$ is a Clifford angle. That is, we have

$$B_{\boldsymbol{\alpha}_c} = U^\dagger(\boldsymbol{\alpha}_c) B U(\boldsymbol{\alpha}_c) \in \mathfrak{p}_n = \{\mathbb{1}, X, Y, Z\}^{\otimes n}. \quad (242)$$

We refer to Lemma 3 for more details and proof.

Now, denote $\text{sup}(B_{\boldsymbol{\alpha}_c})$ as the set of qubits that $B_{\boldsymbol{\alpha}_c}$ acts non-trivially on. The light cone of $B_{\boldsymbol{\alpha}_c}$ is defined as the cardinality of $\text{sup}(B_{\boldsymbol{\alpha}_c})$ i.e., $\Delta_B(\boldsymbol{\alpha}_c) = |\text{sup}(B_{\boldsymbol{\alpha}_c})|$. For example, if $B_{\boldsymbol{\alpha}_c}$ corresponds to $Z \otimes \mathbb{1}^{n-4} \otimes X \otimes Y \otimes Z$, then we have $\Delta_B(\boldsymbol{\alpha}_c) = 4$. Notice the light cone defined this way depends on the Clifford angle $\boldsymbol{\alpha}_c$. We then introduce the concept of an average light cone

$$\Delta_B^{(\text{avg})} = \mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_c} [\Delta_B(\boldsymbol{\alpha}_c)], \quad (243)$$

where $\mathcal{D}_c$ is a uniform distribution over the Clifford parameters. That is, $\mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_c}[\cdot] = \frac{1}{2^M} \sum_{\boldsymbol{\alpha}_c \in \{0, \pi/2\}^M} [\cdot]$. The average light cone intuitively tells us on average how many qubits are involved by the end of the evolution of the circuit $U(\boldsymbol{\alpha})$.

1. Summary of the key technical results

We now present the generic expression of the variance of the MMD loss with general Pauli rotation ansätze $U_{\text{PQC}}(\boldsymbol{\theta})$ for an arbitrary band-width.

Supplemental Proposition 3. *Consider the MMD loss function as defined in Eq. (99) using the Gaussian kernel in Eq. (101) and a QCBM composed of a parametrized Pauli rotation circuit as defined in Eq. (241). Then, we have the variance of the MMD loss scales as*

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] = \text{Var}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})] + 4\text{Var}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\tilde{P}})] - 4\text{Cov}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}}), \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\tilde{P}})] , \quad (244)$$

with

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\tilde{P}})] = \sum_{\substack{A \in \mathcal{N} \\ A \neq \{\}}} (c_{\sigma}(A) z_A(\tilde{P}))^2 \mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_c} \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\boldsymbol{\alpha}_c)} , \quad (245)$$

$$\text{Cov}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}}), \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\tilde{P}})] = 0 , \quad (246)$$

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})] \geq \mathbb{E}_{(A, \boldsymbol{\alpha}_c), (A', \boldsymbol{\alpha}'_c) \sim \tilde{\mathcal{D}}_c(\sigma)} \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\boldsymbol{\alpha}_c) + \Delta_{Z_{A'}}(\boldsymbol{\alpha}'_c)} \left[\left(\frac{9}{5} \right)^{|\text{sup}(Z_{A, \boldsymbol{\alpha}_c}) \cap \text{sup}(Z_{A', \boldsymbol{\alpha}'_c})|} \left(\frac{1}{3} \right)^{n(Z_{A, \boldsymbol{\alpha}_c}, Z_{A', \boldsymbol{\alpha}'_c})} - 1 \right] \geq 0 \quad (247)$$

where $z_A(\tilde{P}) = \text{Tr}[Z_A \rho_{\tilde{P}}]$, $\boldsymbol{\alpha}_c$ are Clifford angles i.e., $\boldsymbol{\alpha}_c \in \{0, \pi/2\}^M$, $\mathcal{D}_c$ is a uniform distribution over these Clifford angles, $Z_{A, \boldsymbol{\alpha}_c} = U^\dagger(\boldsymbol{\alpha}_c) Z_A U(\boldsymbol{\alpha}_c)$ is guaranteed to be a single Pauli-string i.e., $Z_{A, \boldsymbol{\alpha}_c} = \bigotimes_{i=1}^n Z_{A, \boldsymbol{\alpha}_c}^{(i)} \in \{\mathbb{1}, X, Y, Z\}^{\otimes n}$ (which is Z_A back-propagated with $U(\boldsymbol{\alpha}_c)$) together with $\text{sup}(Z_{A, \boldsymbol{\alpha}_c})$ as a set of qubits that $Z_{A, \boldsymbol{\alpha}_c}$ acts non-trivially on and $\Delta_{Z_A}(\boldsymbol{\alpha}_c)$ as the light cone.

In addition, $\tilde{\mathcal{D}}_c(\sigma)$ is the probability distribution that A is chosen with probability $c_{\sigma}(A)$ (from the set $\mathcal{N}$) and independently $\boldsymbol{\alpha}_c$ is uniformly chosen from $\{0, \pi/2\}^M$, as well as $n(Z_{A, \boldsymbol{\alpha}_c}, Z_{A', \boldsymbol{\alpha}'_c})$ is the number of qubits acted by different non-trivial single-qubit Pauli observables $Z_{A, \boldsymbol{\alpha}_c}^{(i)}$ and $Z_{A', \boldsymbol{\alpha}'_c}^{(i)}$ such that $Z_{A, \boldsymbol{\alpha}_c}^{(i)} \neq Z_{A', \boldsymbol{\alpha}'_c}^{(i)} \neq \mathbb{1}$.

We remark that although some proof techniques to compute the variance of the overlap term are similar to Ref. [9], the main technical challenges are to analytically show that the covariance between two terms vanish and compute the exact form of the variance lower bound of the purity term. Now, we recall that for the linear bandwidth $\sigma \in \Theta(n)$ the MMD observable behaves effectively as a few-body observable (see Proposition 4) which allows for some initial trainability guarantee of the QCBM. The formal version of Theorem 3 in the main text is presented below.

Theorem 3 (Pauli rotation ansatz MMD trainability, formal). *Under the same assumptions as in Supplemental Proposition 3 together with the linear bandwidth $\sigma \in \Theta(n)$, as long as there exists some $A \in \mathcal{N}$ with $|A| \in \mathcal{O}(\log(n))$ such that the average light cone of some Z_A back-propagated with $U(\boldsymbol{\alpha}_c)$ as defined in Eq. (243) remains at most in the order of $\log(n)$ qubits, i.e.,*

$$\Delta_{Z_A}^{(\text{avg})} \in \mathcal{O}(\log(n)) , \quad (248)$$

and as long as $z_A^2(\tilde{P})$, which encodes the information about the training distribution, decays at worst polynomially, i.e.,

$$z_A^2(\tilde{P}) \in \Omega(1/\text{poly}(n)) , \quad (249)$$

then the QCBM is trainable in the sense that

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] \in \Omega(1/\text{poly}(n)) . \quad (250)$$

We note that, similar to the discussion at the end of the previous sub-section, the condition on $z_A(\tilde{P})$ is mild and is expected to hold generally in practice as long as few-body marginals of the training distribution does not look too uniform.

The theorem provides a generic criteria in practice for determining when an ansatz does not exhibit a barren plateau. An example of an ansatz that satisfies the few-body light cone condition is a shallow depth circuit with nearest neighbour connectivity (which could be either hardware efficient or problem inspired). Crucially, we emphasise that this light cone argument goes beyond the 2-design assumption and is expected to work even more generally to any ansatz that may not even be in the Pauli rotation form.

Now, we provide details on how to numerically and efficiently estimate this average light cone of the Pauli rotations in practice. For a given Pauli Z_A , we can apply Monte Carlo sampling to estimate the average light cone over the uniform distribution of the Clifford angles. Particularly, we do the following

1. Uniformly sample a Clifford angle α_c ,
2. Work out the light cone of Z_A back-propagated with $U(\alpha_c)$.

This latter step is efficient since for each layer of Pauli-rotation and a fixed Clifford gate it always map a Pauli string to only a single Pauli string (i.e., Lemma 3) and the computational cost of this step scales at worst polynomially with M . With N_{sample} repetitions of steps 1. and 2., the empirical estimate of the average light converges at the rate $\mathcal{O}(1/\sqrt{N_{\text{sample}}})$.

Lastly, to derive the polynomial guarantee on the MMD variance, we simply throw away the purity term which is always guaranteed to be non-negative. However, in practice, we expect the term to contribute to the whole MMD variance and does not generally vanish. Particularly, since the key quantity $z_A^2(\theta)$ of the purity also has the polynomially large variance with the few-body light cone, this intuitively implies the non-vanishing contribution of the purity term to the whole MMD variance. Thus, the theoretical lower bound here is expected to be sub-optimal, yet it is sufficient to show the trainability guarantee.

2. Preliminaries: useful identities

In this subsection, we introduce the following useful mathematical statements. We encourage the readers to only skim through this part and revisit when going through the proof of the main theoretical statements in the next subsections.

Lemma 3. Consider a parametrized ansatz of the form $U(\alpha) = \prod_{k=1}^M e^{-i\alpha_k G_k/2} V_k$ such that the generators $\{G_k\}_{k=1}^M$ are some n -qubit Pauli strings $G_k \in \mathfrak{p}_n = \{\mathbb{1}, X, Y, Z\}^{\otimes n}$ and $\{V_k\}_{k=1}^M$ is a set of non-parametrized Clifford gates, as well as all parameters are uncorrelated. Then, for the Clifford angles $\alpha_c := \alpha \in \{0, \pi/2\}^M$ and a Pauli string B , the effective observable from back propagating $U(\alpha_c)$ remains a single Pauli string (up to a sign factor). That is, we have

$$B_{\alpha_c} = U^\dagger(\alpha_c) B U(\alpha_c) \in \pm \mathfrak{p}_n. \quad (251)$$

Proof. Let us first consider a simpler ansatz of the form

$$U(\alpha) = e^{-i\alpha G/2} V, \quad (252)$$

where the generator is some Pauli-string $G \in \{\mathbb{1}, X, Y, Z\}^{\otimes n}$ and V is some Clifford gate. Then, we have

$$B_\alpha = U^\dagger(\alpha) B U(\alpha) \quad (253)$$

$$= V^\dagger [\cos(\alpha/2)\mathbb{1} + i \sin(\alpha/2)G] B [\cos(\alpha/2)\mathbb{1} - i \sin(\alpha/2)G] V \quad (254)$$

$$= \cos^2(\alpha/2) V^\dagger B V + \sin^2(\alpha/2) V^\dagger G B G V + \sin(\alpha) V^\dagger \left(\frac{i}{2} [G, B] \right) V, \quad (255)$$

where we use the identity $e^{-i\alpha G/2} = \cos(\alpha/2)\mathbb{1} - i \sin(\alpha/2)G$. Since both G and B are Pauli strings, they either commute or anti-commute with one another, which leads to

$$B_\alpha = \begin{cases} V^\dagger B V & \text{if } [B, G] = 0, \\ \cos(\alpha) V^\dagger B V + \sin(\alpha) V^\dagger \left(\frac{i}{2} [G, B] \right) V & \text{if } \{B, G\} = 0. \end{cases} \quad (256)$$

From this, we can see that averaging over α leads to a vanishing expectation (i.e. first order moment vanishes if the Pauli generator of the rotation anti-commutes with the Pauli it acts on). Now, in the case of $\{\sigma, P\} = 0$, we have the following for the Clifford angles i.e., $\alpha \in \{0, \pi/2\}$

$$B_{\alpha_c} = \begin{cases} V^\dagger B V & \text{if } \alpha_c = 0, \\ V^\dagger \left(\frac{i}{2} [G, B] \right) V & \text{if } \alpha_c = \pi/2. \end{cases} \quad (257)$$

Crucially, $\frac{i}{2} [G, B]$ remains a Pauli string. Notice that this is true up to a sign factor, but it will not impact our results in the following. In addition, since V is a Clifford gate, it maps a Pauli-string to another Pauli-string, which implies

$$V^\dagger B V, V^\dagger \left(\frac{i}{2} [G, B] \right) V \in \pm \{\mathbb{1}, X, Y, Z\}^{\otimes n}. \quad (258)$$

That is, for a Clifford angle, the effective observable remains a single Pauli string. We can iteratively apply the result obtained from this simple ansatz to a more general ansatz of the form $U(\alpha) = \prod_{k=1}^M e^{-i\alpha_k G_k/2} V_k$ when the parameter settings are all Clifford angles $\alpha_c := \alpha \in \{0, \pi/2\}^M$, since the effective observable for each layer is a single Pauli string. This completes the proof of the lemma. $\square$

Supplemental Proposition 4. Consider a parametrized ansatz of the form $U(\alpha) = \prod_{k=1}^M e^{-i\alpha_k G_k/2} V_k$ such that the generators $\{G_k\}_{k=1}^M$ are some n -qubit Pauli strings $G_k \in \{\mathbb{1}, X, Y, Z\}^n$ and $\{V_k\}_{k=1}^M$ is a set of non-parametrized Clifford gates, as well as another tensor product ansatz of the form $U_{\text{tensor}}(\beta) = \bigotimes_{i=1}^n U_i(\beta_i)$. Also, consider Pauli strings B and B' which are not identities i.e., $B, B' \subseteq \{\mathbb{1}, X, Y, Z\}^{\otimes n} \setminus \mathbb{1}^{\otimes n}$. Then, if all parameters are uncorrelated, we have the following quantities related to higher moments expressed as

$$\mathbb{E}_\alpha (U^\dagger(\alpha))^{\otimes 2} B^{\otimes 2} (U(\alpha))^{\otimes 2} = \mathbb{E}_{\alpha_c \sim \mathcal{D}_c} (B_{\alpha_c} \otimes B_{\alpha_c}) \quad (259)$$

$$\mathbb{E}_{\alpha, \beta} \text{Tr} \left[(U_{\text{tensor}}(\beta))^{\otimes 2} \rho_0^{\otimes 2} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 2} (U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2} \right] = \delta_{B, B'} \mathbb{E}_{\alpha_c \sim \mathcal{D}_c} \left(\frac{1}{3} \right)^{\Delta_B(\alpha_c)} \quad (260)$$

$$\mathbb{E}_{\alpha, \beta} \text{Tr} \left[(U_{\text{tensor}}(\beta))^{\otimes 3} \rho_0^{\otimes 3} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 3} (U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3} \right] = 0, \quad (261)$$

where $\delta_{B, B'}$ is the Kronecker delta of B and B' , $\alpha_c \in \{0, \pi/2\}^M$ is a Clifford angle, $B_{\alpha_c} = U^\dagger(\alpha_c) B U(\alpha_c)$, $\Delta_B(\alpha_c)$ is a light cone associated with B_{α_c} , and $\mathcal{D}_c$ is a uniform distribution over all possible α_c .

Proof. In the proof, we use a vectorization formalism and refer the readers who are not familiar to, for example, Sec. 5 of Ref. [10].

1. Second moment.

1.1 Simple ansatz. We begin with a simpler ansatz of the form

$$U(\alpha) = e^{-i\alpha G/2} V, \quad (262)$$

where the generator is some Pauli-string $G \in \{\mathbb{1}, X, Y, Z\}^{\otimes n}$ and V is some Clifford gate. Denote the following notation

$$\mathcal{U}_G^{(t)}(\alpha) = \left(e^{i\alpha G/2} \right)^{\otimes t} \otimes \left(e^{-i\alpha G^T/2} \right)^{\otimes t}. \quad (263)$$

The vectorized form of $\mathbb{E}_\alpha (U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2}$ can be expressed as

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2}\rangle = \mathbb{E}_\alpha (U^\dagger(\alpha))^{\otimes 2} \otimes (U^T(\alpha))^{\otimes 2} |B \otimes B'\rangle \quad (264)$$

$$= \mathbb{E}_\alpha \left((V^\dagger)^{\otimes 2} \otimes (V^T)^{\otimes 2} \right) \left(\left(e^{i\alpha G/2} \right)^{\otimes 2} \otimes \left(e^{-i\alpha G^T/2} \right)^{\otimes 2} \right) |B \otimes B'\rangle \quad (265)$$

$$= \left((V^\dagger)^{\otimes 2} \otimes (V^T)^{\otimes 2} \right) \mathbb{E}_\alpha \left[\mathcal{U}_G^{(2)}(\alpha) \right] |B \otimes B'\rangle, \quad (266)$$

where the first equality is due to the transpose-trick $|ABC\rangle = A \otimes C^T |B\rangle$, and the second equality is due to $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and linearity of the expectation. To proceed, there are three different cases to consider together with using $e^{-i\alpha G/2} = \cos(\alpha/2)\mathbb{1} - i\sin(\alpha/2)G$, leading to

1.1.(i) If G does not commute with B and B' i.e. $\{G, B\} = \{G, B'\} = 0$, we have

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2}| = \frac{|V^\dagger B V \otimes V^\dagger B' V\rangle + |V^\dagger (iGB) V \otimes V^\dagger (iGB') V\rangle}{2} \quad (267)$$

$$= \left((V^\dagger)^{\otimes 2} \otimes (V^T)^{\otimes 2} \right) \left(\frac{\mathcal{U}_G^{(2)}(0) + \mathcal{U}_G^{(2)}(\pi/2)}{2} \right) |B \otimes B'\rangle. \quad (268)$$

1.1.(ii) If G commutes with both B and B' i.e., $[G, B] = [G, B'] = 0$, we have

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2}| = |V^\dagger B V \otimes V^\dagger B' V\rangle \quad (269)$$

$$= \left((V^\dagger)^{\otimes 2} \otimes (V^T)^{\otimes 2} \right) \left(\frac{\mathcal{U}_G^{(2)}(0) + \mathcal{U}_G^{(2)}(\pi/2)}{2} \right) |B \otimes B'\rangle, \quad (270)$$

where, to express in the same form in the previous case, we add the action of $\mathcal{U}_G^{(2)}(\pi/2)$ on $|B \otimes B'\rangle$ which simply acts as an identity in this case (due to the commutation assumption).

Crucially, note that iGB and iGB' are simply Pauli-strings i.e., the result of applying $\mathcal{U}_G^{(2)}(\pi/2)$ on to $|B \otimes B'\rangle$ in the 1.1.(i) case. Since V is a Clifford gate which transforms a Pauli-string to another Pauli-string, we have that Eq. (268) results in an equally weighted sum of two Pauli-strings, and Eq. (270) results in a single Pauli-string.

1.1.(iii) If G commutes with only either B or B' e.g., $\{G, B\} = 0$ and $[G, B'] = 0$ (or vice versa), we have

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2}| = \mathbb{E}_\alpha |(U^\dagger(\alpha) \otimes \mathbb{1})(B \otimes B')(U(\alpha) \otimes \mathbb{1})| \quad (271)$$

$$= 0. \quad (272)$$

where the first equality is by the assumption that $[G, B'] = 0$ which simplifies the calculation to compute the first moment with respect to B , resulting in zero when averaging over the full angle range (see Eq. (256)). The same calculation can be done for the case of $[G, B] = 0$ and $\{G, B'\} \neq 0$.

1.2 Generic Pauli ansatz. Now, we extend these findings to an ansatz (with uncorrelated parameters) of the form

$$U(\alpha) U_{\text{tensor}}(\beta) = \left(\prod_{k=1}^M e^{-i\alpha_k G_k/2} V_k \right) \left(\bigotimes_{i=1}^n U_i(\beta_i) \right), \quad (273)$$

such that the generators $\{G_k\}_{k=1}^M$ are some n -qubit Pauli strings $G_k \in \{\mathbb{1}, X, Y, Z\}^n$ and $\{V_k\}_{k=1}^M$ is a set of non-parametrized Clifford gates, as well as all $U_i(\beta_i)$ is a parametrized single-qubit Haar random on the qubit i .

1.2.(i) The scenario when $B = B'$. The vectorized form of $\mathbb{E}_\alpha (U^\dagger(\alpha))^{\otimes 2} (B \otimes B) (U(\alpha))^{\otimes 2}$ can be expressed as

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 2} (B \otimes B) (U(\alpha))^{\otimes 2}\rangle = \mathbb{E}_\alpha [(U^\dagger(\alpha))^{\otimes 2} \otimes (U^T(\alpha))^{\otimes 2}] |B \otimes B\rangle \quad (274)$$

$$= \mathbb{E}_\alpha \left[\prod_{k=M}^1 \left((V_k^\dagger)^{\otimes 2} \otimes (V_k^T)^{\otimes 2} \right) \left((e^{i\alpha_k G_k/2})^{\otimes 2} \otimes (e^{-i\alpha_k G_k^T/2})^{\otimes 2} \right) \right] |B \otimes B\rangle \quad (275)$$

$$= \prod_{k=M}^1 (V_k^\dagger)^{\otimes 2} \otimes (V_k^T)^{\otimes 2} \mathbb{E}_{\alpha_k} [\mathcal{U}_{G_k}^{(2)}(\alpha_k)] |B \otimes B\rangle \quad (276)$$

$$= \left(\prod_{k=M-1}^1 (V_k^\dagger)^{\otimes 2} \otimes (V_k^T)^{\otimes 2} \mathbb{E}_{\alpha_k} [\mathcal{U}_{G_k}^{(2)}(\alpha_k)] \right) \cdot \left(\frac{\mathcal{U}_{G_M}^{(2)}(0) + \mathcal{U}_{G_M}^{(2)}(\pi/2)}{2} \right) |B \otimes B\rangle \quad (277)$$

$$= \frac{1}{2^M} \sum_{\alpha_c \in \{0, \pi/2\}^M} \left[\prod_{k=M}^1 (V_k^\dagger)^{\otimes 2} \otimes (V_k^T)^{\otimes 2} \mathcal{U}_{G_k}^{(2)}(\alpha_c^{(k)}) \right] |B \otimes B\rangle \quad (278)$$

$$= \frac{1}{2^M} \sum_{\alpha_c \in \{0, \pi/2\}^M} |B_{\alpha_c} \otimes B_{\alpha_c}\rangle \quad (279)$$

$$= \mathbb{E}_{\alpha_c \sim \mathcal{D}_c} |B_{\alpha_c} \otimes B_{\alpha_c}\rangle \quad (280)$$

where in the second line we introduce the notation $\prod_{k=M}^1 a_k = a_1 a_2 \dots a_{M-1} a_M$, the third equality is due to parameters being uncorrelated and by using the notation $\mathcal{U}_{G_k}^{(2)}(\alpha_k)$ introduced in Eq. (263). The fourth equality is by either Eq. (268) if $\{B, G_M\} = 0$ or Eq. (270) if $[B, G_M] = 0$, which happily results in the same expression. The fifth equality is the result of recursively applying what we just do for all layers. Crucially, thanks to Lemma 3, each term in the sum correspond to two copies of a single Pauli string. In the last equality, we simply use the definition of uniformly average over the Clifford angles i.e., $\mathbb{E}_{\alpha_c \sim \mathcal{D}_c}[\cdot] = \frac{1}{2^M} \sum_{\alpha_c \in \{0, \pi/2\}^M} [\cdot]$. We note that this proves Eq. (259) in the supplemental proposition.

Now, we can do the average over the final tensor-product layer $U_{\text{tensor}}(\beta) = \bigotimes_{i=1}^n U_i(\beta_i)$ by expressing $B_{\alpha_c} = \bigotimes_{i=1}^n B_{\alpha_c}^{(i)}$ with $B_{\alpha_c}^{(i)} \in \{\pm \mathbb{1}, X, Y, Z\}$ and performing Haar integration qubit-wise, leading to

$$\mathbb{E}_{\alpha, \beta} \text{Tr} [(U_{\text{tensor}}(\beta))^{\otimes 2} \rho_0^{\otimes 2} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 2} (U^\dagger(\alpha))^{\otimes 2} B^{\otimes 2} (U(\alpha))^{\otimes 2}] = \mathbb{E}_\beta \text{Tr} [(U_{\text{tensor}}(\beta))^{\otimes 2} \rho_0^{\otimes 2} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 2} \mathbb{E}_{\alpha_c \sim \mathcal{D}_c} (B_{\alpha_c}^{\otimes 2})] \quad (281)$$

$$= \mathbb{E}_{\alpha_c \sim \mathcal{D}_c} \prod_{i=1}^n \mathbb{E}_{\beta_i} \text{Tr} [(U_i(\beta_i))^{\otimes 2} \rho_0^{(i)} (U_i^\dagger(\beta_i))^{\otimes 2} B_{\alpha_c}^{(i)\otimes 2}] \quad (282)$$

$$= \mathbb{E}_{\alpha_c \sim \mathcal{D}_c} \left(\frac{1}{3} \right)^{\Delta_B(\alpha_c)}, \quad (283)$$

where in the last equality, the Haar integration from Eq. (153) gives 1/3 if $B_{\alpha_c}^{(i)} \neq \mathbb{1}$ (otherwise, it gives 1).

1.2.(ii) The scenario when $B \neq B'$. For two initial Pauli strings B and B' such that $B \neq B'$, they evolve into two different Pauli-strings B_{α_c} and B'_{α_c} with the same $U(\alpha_c)$. This can be trivially seen since $B_{\alpha_c} = U^\dagger(\alpha_c) B U(\alpha_c)$ and $B'_{\alpha_c} = U^\dagger(\alpha_c) B' U(\alpha_c)$ they can evolve to the same Pauli-string only if $B = B'$. Hence, the average over α in the vectorized form can be expressed as

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2}\rangle = \frac{1}{2^M} \sum_{\alpha_c \in \{0, \pi/2\}^M} c_{B, B'}(\alpha_c) |B_{\alpha_c} \otimes B'_{\alpha_c}\rangle \quad (284)$$

with $B_{\alpha_c} \neq B'_{\alpha_c}$. The coefficient $c_{B, B'}(\alpha_c)$ takes into account the fact that we could have some generator that commutes with only one of the two evolving Pauli-string (i.e., the 1.1.(iii) case described in the simple ansatz scenario above). $c_{B, B'}(\alpha_c)$ takes the value of either 1 if there is no generator that commutes with only one evolving Pauli-string, or 0 if there exists a generator that commutes with only one evolving Pauli-string (due to Eq. (272)).

Crucially, averaging over the tensor-product ansatz $U_{\text{tensor}}(\beta)$ ensures that the contribution of any pair of different Pauli-strings vanishes. That is, for any two different Pauli-strings B_1 and B_2 such that $B_1 \neq B_2$, we have

$$\mathbb{E}_\beta \text{Tr} \left[(U_{\text{tensor}}(\beta))^{\otimes 2} \rho_0^{\otimes 2} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 2} (B_1 \otimes B_2) \right] = \prod_{i=1}^n \mathbb{E}_{\beta_i} \text{Tr} \left[(U_i(\beta_i))^{\otimes 2} \rho_0^{(i)\otimes 2} (U_i^\dagger(\beta_i))^{\otimes 2} (B_1^{(i)} \otimes B_2^{(i)}) \right] \quad (285)$$

$$= 0, \quad (286)$$

which is from the direct computation of Haar integration with Eq. (153).

From the 1.2.(ii) case, we have that for different initial Pauli-strings B and B' we have

$$\mathbb{E}_{\alpha, \beta} \text{Tr} \left[(U_{\text{tensor}}(\beta))^{\otimes 2} \rho_0^{\otimes 2} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 2} (U^\dagger(\alpha))^{\otimes 2} (B \otimes B') (U(\alpha))^{\otimes 2} \right] = 0. \quad (287)$$

Together from the 1.2.(i) and 1.2.(ii) cases, the proof of Eq. (260) is complete.

2. Third moment. We take the same strategy as in the second moment.

2.1 Simple ansatz. We consider the same simple ansatz of the form in Eq. (262). Then, the vectorized form of $\mathbb{E}_\alpha (U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3}$ can be expressed as

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3}\rangle = \mathbb{E}_\alpha (U^\dagger(\alpha))^{\otimes 3} \otimes (U^T(\alpha))^{\otimes 3} |B^{\otimes 2} \otimes B'\rangle \quad (288)$$

$$= \mathbb{E}_\alpha \left((V^\dagger)^{\otimes 3} \otimes (V^T)^{\otimes 3} \right) \left((e^{i\alpha G/2})^{\otimes 3} \otimes (e^{-i\alpha G^T/2})^{\otimes 3} \right) |B^{\otimes 2} \otimes B'\rangle \quad (289)$$

$$= \left((V^\dagger)^{\otimes 3} \otimes (V^T)^{\otimes 3} \right) \mathbb{E}_\alpha [\mathcal{U}_G^{(3)}(\alpha)] |B^{\otimes 2} \otimes B'\rangle, \quad (290)$$

where the first equality is due to the transpose-trick $|ABC\rangle = A \otimes C^T |B\rangle$, and the second equality is due to $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and linearity of the expectation. To proceed, there are three different cases to consider together with using $e^{-i\alpha G/2} = \cos(\alpha/2)\mathbb{1} - i\sin(\alpha/2)G$, which lead to

2.1.(i) If G does not commute with B' i.e. $\{G, B'\} \neq 0$, we further have two sub-scenarios including (i).a for $[G, B] = 0$ and (i).b for $\{G, B\} = 0$. Either sub-scenario results in zero when averaging over i.e.,

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3}\rangle = 0. \quad (291)$$

Particularly, in the (i).a sub-scenario with $[G, B] = 0$, the computation reduces to compute the first moment which results in zero. On the other hand, in the (i).b sub-scenario with $\{G, B\} = 0$, we have

$$\begin{aligned} \mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3}\rangle &= \mathbb{E}_\alpha [\cos^3(\alpha)] |B^{\otimes 2} \otimes B'\rangle + \mathbb{E}_\alpha [\sin^2(\alpha) \cos(\alpha)] (iGB)^{\otimes 2} \otimes B'\rangle \\ &\quad + \mathbb{E}_\alpha [\cos^2(\alpha) \sin(\alpha)] (|B \otimes iGB \otimes B'\rangle + |iGB \otimes B \otimes B'\rangle) \\ &\quad + \mathbb{E}_\alpha [\cos^2(\alpha) \sin(\alpha)] |B^{\otimes 2} \otimes iGB\rangle + \mathbb{E}_\alpha [\sin^3(\alpha)] (iGB)^{\otimes 2} \otimes iGB'\rangle \\ &\quad + \mathbb{E}_\alpha [\cos(\alpha) \sin^2(\alpha)] (|B \otimes iGB \otimes iGB'\rangle + |iGB \otimes B \otimes iGB'\rangle) \\ &= 0. \end{aligned} \quad (292)$$

Indeed, this (i).b sub-scenario is the only scenario that is of particular interest regarding the third moment. Other scenarios reduce themselves to either compute the first or second moments over the parameters α , as we already saw in the (i).a sub-scenario.

2.1.(ii) If G commutes with B' and does not commute with B i.e., $[G, B'] = 0$ and $\{G, B\} \neq 0$, this reduces to compute the second moment (see the case 1.1.(i) above i.e., Eq. (268)), leading to

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3}\rangle = \left((V^\dagger)^{\otimes 2} \otimes (V^T)^{\otimes 2} \right) \left(\frac{\mathcal{U}_G^{(2)}(0) + \mathcal{U}_G^{(2)}(\pi/2)}{2} \right) |B \otimes B\rangle \otimes (V^\dagger \otimes V^T) |B'\rangle. \quad (294)$$

2.1.(iii) If G commutes with B' and commutes with B i.e., $[G, B'] = 0$ and $[G, B] = 0$, this reduces to the parameterised part of the circuit completely commuting with the observables, left them unchanged. The result can also be written in the same form as in the previous case (see the case 1.1.(ii) above i.e., Eq. (270))

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3} \rangle = (V^\dagger)^{\otimes 3} \otimes (V^T)^{\otimes 3} |B^{\otimes 2} \otimes B'\rangle \quad (295)$$

$$= \left((V^\dagger)^{\otimes 2} \otimes (V^T)^{\otimes 2} \right) \left(\frac{\mathcal{U}_G^{(2)}(0) + \mathcal{U}_G^{(2)}(\pi/2)}{2} \right) |B \otimes B\rangle \otimes (V^\dagger \otimes V^T) |B'\rangle, \quad (296)$$

where $\mathcal{U}_G^{(2)}(\pi/2)$ simply commutes through and effectively acts as an identity.

2.2 Generic Pauli ansatz. Now, we generalises our findings to Pauli rotation ansätze (with uncorrelated parameters) of the form in Eq. (273).

To analytically show that the third moment vanishes, we can keep applying the results in the simple ansatz scenario to each layer. We notice that if the 2.1.(i) case in the simple ansatz above happens the third moment of the generic Pauli rotation ansatz simply vanishes. For example, this happens when B' does not commute with G_M . More generally, if there exists at least one generator that does not commute with the evolving Pauli-string with the initial B' at that particular layer, then this implies the 2.1.(i) case happens.

In the event that we never satisfy the 2.1.(i) case this means the evolving B' always commutes with the generator. In other words, the evolving B' does not contribute at all to the third moment regarding $U(\alpha)$. By repeatedly applying the 2.1.(ii) and 2.1.(iii) cases, the evolution of $B^{\otimes 2}$ can be obtained to be identical to the 1.2.(i) case in the second moment above. Thus, the vectorized form of the average over $U(\alpha)$ is

$$\mathbb{E}_\alpha |(U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3} \rangle = \mathbb{E}_{\alpha_c \sim \mathcal{D}_c} |B_{\alpha_c} \otimes B_{\alpha_c}\rangle \otimes |B'_V\rangle, \quad (297)$$

with $|B'_V\rangle = \prod_{k=M}^1 (V_k^\dagger \otimes V_k^T) |B'\rangle$ which is still a single Pauli-string due to the Clifford gates $\{V_k\}_{k=1}^M$.

Now, the average over the tensor-product ansatz $U_{\text{tensor}}(\beta)$ ensures the third moment disappears. Since for any two different Pauli-strings B_1 and B_2 that are not identity such that $B_1 \neq B_2$, we have

$$\mathbb{E}_\beta \text{Tr} \left[(U_{\text{tensor}}(\beta))^{\otimes 3} \rho_0^{\otimes 3} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 3} (B_1^{\otimes 2} \otimes B_2) \right] = \prod_{i=1}^n \mathbb{E}_{\beta_i} \text{Tr} \left[U_i(\beta_i)^{\otimes 3} \rho_0^{(i)\otimes 3} U_i^{\dagger\otimes 3}(\beta_i) (B_1^{(i)\otimes 2} \otimes B_2^{(i)}) \right] \quad (298)$$

$$= 0, \quad (299)$$

which is due to the direct computation of Haar integration using Eq. (154). Due to the linearity of average and by applying Eq. (299) to each term in Eq. (297), we have the third moment to vanish in this scenario.

Hence, altogether the third moment vanishes in all scenarios

$$\mathbb{E}_{\alpha, \beta} \text{Tr} \left[(U_{\text{tensor}}(\beta))^{\otimes 3} \rho_0^{\otimes 3} (U_{\text{tensor}}^\dagger(\beta))^{\otimes 3} (U^\dagger(\alpha))^{\otimes 3} (B^{\otimes 2} \otimes B') (U(\alpha))^{\otimes 3} \right] = 0. \quad (300)$$

This completes the proof of the supplemental proposition. $\square$

Lemma 4. Consider a function $f(\theta) := f(\alpha, \beta)$ with $\theta := (\alpha, \beta)$. The variance of the function can be decomposed as

$$\text{Var}_\theta[f(\theta)] = \mathbb{E}_\alpha[\text{Var}_\beta f(\alpha, \beta)] + \text{Var}_\alpha[\mathbb{E}_\beta f(\alpha, \beta)]. \quad (301)$$

Proof. We can express the variance of $f(\theta)$ as

$$\text{Var}_\theta[f(\theta)] = \text{Var}_{\alpha, \beta}[f(\alpha, \beta)] \quad (302)$$

$$= \mathbb{E}_{\alpha, \beta}[f^2(\alpha, \beta)] - (\mathbb{E}_{\alpha, \beta}[f(\alpha, \beta)])^2 \quad (303)$$

$$= \mathbb{E}_{\alpha, \beta}[f^2(\alpha, \beta)] + \left(\mathbb{E}_\alpha \left[\mathbb{E}_\beta [f(\alpha, \beta)]^2 \right] - \mathbb{E}_\alpha \left[\mathbb{E}_\beta [f(\alpha, \beta)]^2 \right] \right) - (\mathbb{E}_{\alpha, \beta}[f(\alpha, \beta)])^2 \quad (304)$$

$$= \mathbb{E}_\alpha \left(\mathbb{E}_\beta [f^2(\alpha, \beta)] - \mathbb{E}_\beta [f(\alpha, \beta)]^2 \right) + \left(\mathbb{E}_\alpha \left[\mathbb{E}_\beta [f(\alpha, \beta)]^2 \right] - (\mathbb{E}_\alpha [\mathbb{E}_\beta f(\alpha, \beta)])^2 \right) \quad (305)$$

$$= \mathbb{E}_\alpha [\text{Var}_\beta f(\alpha, \beta)] + \text{Var}_\alpha [\mathbb{E}_\beta f(\alpha, \beta)], \quad (306)$$

where in the third equality we add and subtract $\mathbb{E}_{\alpha} [\mathbb{E}_{\beta} [f(\alpha, \beta)]^2]$ and to reach the last equality we simply use the definition of the variance. This completes the proof of the lemma. $\square$

3. Proof of Supplemental Proposition 3

Proof. The variance of the MMD loss in Eq. (240) can be written as

$$\text{Var}_{\theta} [\mathcal{L}_{\text{MMD}}(\theta)] = \text{Var}_{\theta} [\mathcal{M}(\rho_{\theta}, \rho_{\theta}) - 2\mathcal{M}(\rho_{\theta}, \rho_{\bar{p}}) + \mathcal{M}(\rho_{\bar{p}}, \rho_{\bar{p}})] \quad (307)$$

$$= \text{Var}_{\theta} [\mathcal{M}(\rho_{\theta}, \rho_{\theta})] + 4\text{Var}_{\theta} [\mathcal{M}(\rho_{\theta}, \rho_{\bar{p}})] - 4\text{Cov}_{\theta} [\mathcal{M}(\rho_{\theta}, \rho_{\theta}), \mathcal{M}(\rho_{\theta}, \rho_{\bar{p}})], \quad (308)$$

where we have used $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$ and $\text{Var}[X+c] = \text{Var}[X]$ for any random variables X, Y and some constant c .

Our strategy is to tackle Eq. (308) term by term.

(i) the variance of the cross term $\mathcal{M}(\rho_{\theta}, \rho_{\bar{p}})$. The cross term can be expressed as

$$\mathcal{M}(\rho_{\theta}, \rho_{\bar{p}}) = \text{Tr} [O_{\text{MMD}}^{(\sigma)}(\rho_{\theta} \otimes \rho_{\bar{p}})] \quad (309)$$

$$= \text{Tr} \left[\sum_{A \in \mathcal{N}} c_{\sigma}(A) Z_A \otimes Z_A (\rho_{\theta} \otimes \rho_{\bar{p}}) \right] \quad (310)$$

$$= \sum_{A \in \mathcal{N}} c_{\sigma}(A) z_A(\tilde{P}) z_A(\theta), \quad (311)$$

where $z_A(\tilde{P}) = \text{Tr}[\rho_{\bar{p}} Z_A]$ and $z_A(\theta) = \text{Tr}[U_{\text{PQC}}(\theta) \rho_0 U_{\text{PQC}}^{\dagger}(\theta) Z_A]$. Then, we can compute the variance as

$$\text{Var}_{\theta} [\mathcal{M}(\rho_{\theta}, \rho_{\bar{p}})] = \text{Var}_{\theta} \left[\sum_{A \in \mathcal{N}} c_{\sigma}(A) z_A(\tilde{P}) z_A(\theta) \right] \quad (312)$$

$$= \sum_{A \in \mathcal{N}} (c_{\sigma}(A) z_A(\tilde{P}))^2 \text{Var}_{\theta} [z_A(\theta)] + \sum_{\substack{A, A' \in \mathcal{N} \\ A \neq A'}} c_{\sigma}(A) c_{\sigma}(A') z_A(\tilde{P}) z_{A'}(\tilde{P}) \text{Cov}_{\theta} [z_A(\theta), z_{A'}(\theta)], \quad (313)$$

which is due to a standard formula of variance of a sum i.e., $\text{Var}[\sum_i X_i] = \sum_{i,j} \text{Cov}[X_i, X_j]$. Notice that the identity term, $A = \{ \}$, does not contribute to the variance as it is a constant. Indeed, in this case we have $z_A(\theta) = 1$. So, we are only considering $A, A' \neq \{ \}$ in the following (i.e. we can remove cases where $Z_A = \mathbb{1}$ or $Z_{A'} = \mathbb{1}$ in previous equation). First, we consider the average term

$$\mathbb{E}_{\theta} [z_A(\theta)] = \mathbb{E}_{\alpha, \beta} \text{Tr} [U_{\text{tensor}}(\beta) \rho_0 U_{\text{tensor}}^{\dagger}(\beta) (U^{\dagger}(\alpha) Z_A U(\alpha))] \quad (314)$$

$$= \mathbb{E}_{\beta} \text{Tr} [U_{\text{tensor}}(\beta) \rho_0 U_{\text{tensor}}^{\dagger}(\beta) \mathbb{E}_{\alpha} (U^{\dagger}(\alpha) Z_A U(\alpha))] \quad (315)$$

$$= 0, \quad (316)$$

where we average over the full angle range and hence the first moment disappears in this case.

Next, we consider the covariance term when $A \neq A'$

$$\text{Cov}_{\theta} [z_A(\theta), z_{A'}(\theta)] = \mathbb{E}_{\theta} [z_A(\theta) z_{A'}(\theta)] \quad (317)$$

$$= \mathbb{E}_{\alpha, \beta} \text{Tr} [(U_{\text{tensor}}(\beta))^{\otimes 2} \rho_0^{\otimes 2} (U_{\text{tensor}}^{\dagger}(\beta))^{\otimes 2} (U^{\dagger}(\alpha))^{\otimes 2} (Z_A \otimes Z_{A'}) (U(\alpha))^{\otimes 2}] \quad (318)$$

$$= 0, \quad (319)$$

where to reach the last equality we apply Eq. (260) from Supplemental Proposition 4.

Lastly, we compute the variance terms.

$$\text{Var}_{\boldsymbol{\theta}}[z_A(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta})] \quad (320)$$

$$= \mathbb{E}_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \text{Tr}[(U_{\text{tensor}}(\boldsymbol{\beta}))^{\otimes 2} \rho_0^{\otimes 2} (U_{\text{tensor}}^\dagger(\boldsymbol{\beta}))^{\otimes 2} (U^\dagger(\boldsymbol{\alpha}))^{\otimes 2} Z_A^{\otimes 2} (U(\boldsymbol{\alpha}))^{\otimes 2}] \quad (321)$$

$$= \mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_c} \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\boldsymbol{\alpha}_c)}, \quad (322)$$

where similarly to reach the last equality we apply Eq. (260) from Supplemental Proposition 4.

By substituting these back into Eq. (313), we obtain the variance of the cross term as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\tilde{p}})] = \sum_{\substack{A \in \mathcal{N} \\ A \neq \{\}}} (c_{\sigma}(A) z_A(\tilde{P}))^2 \mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_M} \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\boldsymbol{\alpha}_c)}. \quad (323)$$

(ii) The covariance between the purity term $\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})$ and the cross term $\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\tilde{p}})$ is of the form

$$\text{Cov}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}}), \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\tilde{p}})] = \text{Cov}_{\boldsymbol{\theta}} \left[\sum_{A \in \mathcal{N}} c_{\sigma}(A) z_A^2(\boldsymbol{\theta}), \sum_{A' \in \mathcal{N}} c_{\sigma}(A') z_{A'}(\tilde{P}) z_{A'}(\boldsymbol{\theta}) \right] \quad (324)$$

$$= \sum_{A, A' \in \mathcal{N}} c_{\sigma}(A) c_{\sigma}(A') z_{A'}(\tilde{P}) \text{Cov}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}), z_{A'}(\boldsymbol{\theta})], \quad (325)$$

There are 4 different cases to consider. We note that when $Z_A = Z_{A'} = \mathbb{1}$ we have that $z_A(\boldsymbol{\theta}) = z_{A'}(\boldsymbol{\theta}) = \text{Tr}[U_{\text{PQC}}(\boldsymbol{\theta}) \rho_0 U^\dagger(\boldsymbol{\theta}) \mathbb{1}] = 1$.

For $Z_A \neq \mathbb{1}$ and $Z_{A'} \neq \mathbb{1}$, we have

$$\text{Cov}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}), z_{A'}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}) z_{A'}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta})] \mathbb{E}_{\boldsymbol{\theta}}[z_{A'}(\boldsymbol{\theta})] \quad (326)$$

$$= \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}) z_{A'}(\boldsymbol{\theta})] \quad (327)$$

$$= \mathbb{E}_{\boldsymbol{\theta}} \text{Tr}[(U_{\text{tensor}}(\boldsymbol{\beta}))^{\otimes 3} \rho_0^{\otimes 3} (U_{\text{tensor}}^\dagger(\boldsymbol{\beta}))^{\otimes 3} (U^\dagger(\boldsymbol{\alpha}))^{\otimes 3} (Z_A^{\otimes 2} \otimes Z_{A'}) (U(\boldsymbol{\alpha}))^{\otimes 3}] \quad (328)$$

$$= \mathbb{E}_{\boldsymbol{\beta}} \text{Tr}[(U_{\text{tensor}}(\boldsymbol{\beta}))^{\otimes 3} \rho_0^{\otimes 3} (U_{\text{tensor}}^\dagger(\boldsymbol{\beta}))^{\otimes 3} \mathbb{E}_{\boldsymbol{\alpha}}((U^\dagger(\boldsymbol{\alpha}))^{\otimes 3} (Z_A^{\otimes 2} \otimes Z_{A'}) (U(\boldsymbol{\alpha}))^{\otimes 3})] \quad (329)$$

$$= 0, \quad (330)$$

where the last equality is due to Eq. (261) from Supplemental Proposition 4.

For $Z_A = \mathbb{1}$ and $Z_{A'} \neq \mathbb{1}$, we have

$$\text{Cov}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}), z_{A'}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}) z_{A'}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta})] \mathbb{E}_{\boldsymbol{\theta}}[z_{A'}(\boldsymbol{\theta})] \quad (331)$$

$$= \mathbb{E}_{\boldsymbol{\theta}}[z_{A'}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[z_{A'}(\boldsymbol{\theta})] \quad (332)$$

$$= 0. \quad (333)$$

For $Z_A \neq \mathbb{1}$ and $Z_{A'} = \mathbb{1}$, we have

$$\text{Cov}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}), z_{A'}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}) z_{A'}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta})] \mathbb{E}_{\boldsymbol{\theta}}[z_{A'}(\boldsymbol{\theta})] \quad (334)$$

$$= \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta})] \quad (335)$$

$$= 0. \quad (336)$$

For $Z_A = \mathbb{1}$ and $Z_{A'} = \mathbb{1}$, we have

$$\text{Cov}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}), z_{A'}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta}) z_{A'}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[z_A^2(\boldsymbol{\theta})] \mathbb{E}_{\boldsymbol{\theta}}[z_{A'}(\boldsymbol{\theta})] \quad (337)$$

$$= 1 - 1 \quad (338)$$

$$= 0. \quad (339)$$

Hence, altogether we conclude in all cases that the covariance between the cross term and the purity term vanishes

$$\text{Cov}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}}), \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\bar{\boldsymbol{\theta}}})] = 0. \quad (340)$$

(iii) The variance of the purity term $\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})$. We remind that $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and apply Lemma 4 to obtain the following lower bound

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})] = \text{Var}_{\boldsymbol{\beta}}[\mathbb{E}_{\boldsymbol{\alpha}} \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})] + \mathbb{E}_{\boldsymbol{\beta}}[\text{Var}_{\boldsymbol{\alpha}} \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})] \quad (341)$$

$$\geq \text{Var}_{\boldsymbol{\beta}}[\mathbb{E}_{\boldsymbol{\alpha}} \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})], \quad (342)$$

where we throw away the second non-negative term.

Now, we consider the central object $\mathbb{E}_{\boldsymbol{\alpha}} \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})$ which can be expressed as

$$\mathbb{E}_{\boldsymbol{\alpha}} \mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\alpha}} \sum_{A \in \mathcal{N}} c_{\sigma}(A) (z_A(\boldsymbol{\theta}))^2 \quad (343)$$

$$= \sum_{A \in \mathcal{N}} c_{\sigma}(A) \text{Tr} [U_{\text{tensor}}^{\otimes 2}(\boldsymbol{\beta}) \rho_0^{\otimes 2} U_{\text{tensor}}^{\dagger \otimes 2}(\boldsymbol{\beta}) \mathbb{E}_{\boldsymbol{\alpha}} (U^{\dagger \otimes 2}(\boldsymbol{\alpha}) Z_A^{\otimes 2} U^{\otimes 2}(\boldsymbol{\alpha}))] \quad (344)$$

$$= \sum_{A \in \mathcal{N}} c_{\sigma}(A) \text{Tr} [U_{\text{tensor}}^{\otimes 2}(\boldsymbol{\beta}) \rho_0^{\otimes 2} U_{\text{tensor}}^{\dagger \otimes 2}(\boldsymbol{\beta}) \mathbb{E}_{\boldsymbol{\alpha}_c \sim \tilde{\mathcal{D}}_c} [Z_{A, \boldsymbol{\alpha}_c} \otimes Z_{A, \boldsymbol{\alpha}_c}]] \quad (345)$$

$$= \sum_{A \in \mathcal{N}} c_{\sigma}(A) \mathbb{E}_{\boldsymbol{\alpha}_c \sim \tilde{\mathcal{D}}_c} \text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{\otimes 2}] \quad (346)$$

$$= \mathbb{E}_{(A, \boldsymbol{\alpha}_c) \sim \tilde{\mathcal{D}}_c(\sigma)} \text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{\otimes 2}], \quad (347)$$

where the third equality is due to Eq. (259) in Supplemental Proposition 4 and in the fourth equality we pull out the expectation and introduce the shorthand $\rho_{\boldsymbol{\beta}} = U_{\text{tensor}}(\boldsymbol{\beta}) \rho_0 U_{\text{tensor}}^{\dagger}(\boldsymbol{\beta})$. To reach the last equality, we notice that $c_{\sigma}(A) = (1 - p_{\sigma})^{n-|A|} p_{\sigma}^{|A|}$ (see e.g., Eq. (238)) can be interpreted as a probability of sampling A , and $\sum_{A \in \mathcal{N}} c_{\sigma}(A) = 1$. Hence, $\sum_{A \in \mathcal{N}} c_{\sigma}(A) \mathbb{E}_{\boldsymbol{\alpha}_c \sim \tilde{\mathcal{D}}_c}[\cdot] = \mathbb{E}_{(A, \boldsymbol{\alpha}_c) \sim \tilde{\mathcal{D}}_c(\sigma)}$ where $\tilde{\mathcal{D}}_c(\sigma)$ is the probability distribution that A is chosen with probability $c_{\sigma}(A)$ (from the set $\mathcal{N}$) and independently $\boldsymbol{\alpha}_c$ is uniformly chosen from $\{0, \pi/2\}^M$.

By plugging Eq. (347) back into Eq. (342), we have

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{M}(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}})] \geq \text{Var}_{\boldsymbol{\beta}}[\mathbb{E}_{(A, \boldsymbol{\alpha}_c) \sim \tilde{\mathcal{D}}_c(\sigma)} \text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{\otimes 2}]] \quad (348)$$

$$= \mathbb{E}_{(A, \boldsymbol{\alpha}_c), (A', \boldsymbol{\alpha}'_c) \sim \tilde{\mathcal{D}}_c(\sigma)} \text{Cov}_{\boldsymbol{\beta}}[\text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{\otimes 2}], \text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 2} Z_{A', \boldsymbol{\alpha}'_c}^{\otimes 2}]] \quad (349)$$

$$= \mathbb{E}_{(A, \boldsymbol{\alpha}_c), (A', \boldsymbol{\alpha}'_c) \sim \tilde{\mathcal{D}}_c(\sigma)} [\mathbb{E}_{\boldsymbol{\beta}} \text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 4} (Z_{A, \boldsymbol{\alpha}_c}^{\otimes 2} \otimes Z_{A', \boldsymbol{\alpha}'_c}^{\otimes 2})] - \mathbb{E}_{\boldsymbol{\beta}} \text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{\otimes 2}] \mathbb{E}_{\boldsymbol{\beta}} \text{Tr} [\rho_{\boldsymbol{\beta}}^{\otimes 2} Z_{A', \boldsymbol{\alpha}'_c}^{\otimes 2}]] \quad (350)$$

$$= \mathbb{E}_{(A, \boldsymbol{\alpha}_c), (A', \boldsymbol{\alpha}'_c) \sim \tilde{\mathcal{D}}_c(\sigma)} \left[\prod_{i=1}^n \mathbb{E}_{\beta_i} \text{Tr} [\rho_{\beta_i}^{\otimes 4} (Z_{A, \boldsymbol{\alpha}_c}^{(i)\otimes 2} \otimes Z_{A', \boldsymbol{\alpha}'_c}^{(i)\otimes 2})] - \prod_{i=1}^n \mathbb{E}_{\beta_i} \text{Tr} [\rho_{\beta_i}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{(i)\otimes 2}] \mathbb{E}_{\beta_i} \text{Tr} [\rho_{\beta_i}^{\otimes 2} Z_{A', \boldsymbol{\alpha}'_c}^{(i)\otimes 2}] \right], \quad (351)$$

where the second line is due to $\text{Var}[\sum_i X_i] = \sum_{i,j} \text{Cov}[X_i, X_j]$.

To further proceed, we have to carry out these Haar integration in Eq. (351).

For the second order moment, we remind that the following relation holds through the Haar integration.

$$\mathbb{E}_{\beta_i} \text{Tr} [\rho_{\beta_i}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{(i)\otimes 2}] = \begin{cases} 1/3 & \text{if } Z_{A, \boldsymbol{\alpha}_c}^{(i)} \neq \mathbb{1}, \\ 1 & \text{if } Z_{A, \boldsymbol{\alpha}_c}^{(i)} = \mathbb{1}. \end{cases} \quad (352)$$

For the fourth order moment, we first notice that if either $Z_{A, \boldsymbol{\alpha}_c}^{(i)} = \mathbb{1}$ or $Z_{A', \boldsymbol{\alpha}'_c}^{(i)} = \mathbb{1}$, we do not need to compute 4-design integration. Indeed, we have that $\text{Tr} [\rho_{\beta_i}^{\otimes 4} Z_{A, \boldsymbol{\alpha}_c}^{(i)\otimes 2} \otimes \mathbb{1}^{\otimes 2}] = \text{Tr} [\rho_{\beta_i}^{\otimes 2} Z_{A, \boldsymbol{\alpha}_c}^{(i)\otimes 2}]$. Now, for the cases where neither $Z_{A, \boldsymbol{\alpha}_c}^{(i)}$ nor

$Z_{A',\alpha'_c}^{(i)}$ is $\mathbb{1}$, we use Eq. (155) which leads to

$$\mathbb{E}_{\beta_i} \text{Tr} \left[\rho_{\beta_i}^{\otimes 4} (Z_{A,\alpha_c}^{(i)\otimes 2} \otimes Z_{A',\alpha'_c}^{(i)}) \right] = \begin{cases} 1/5 & \text{if } Z_{A,\alpha_c}^{(i)} = Z_{A',\alpha'_c}^{(i)}, \\ 1/15 & \text{if } Z_{A,\alpha_c}^{(i)} \neq Z_{A',\alpha'_c}^{(i)}. \end{cases} \quad (353)$$

Given Pauli-strings $Z_{A,\alpha_c} = \bigotimes_{i=1}^n Z_{A,\alpha_c}^{(i)}$ and $Z_{A',\alpha'_c} = \bigotimes_{i=1}^n Z_{A',\alpha'_c}^{(i)}$, we do a qubit-wise comparison to see which case for each qubit in these following four different cases it is in

1. $Z_{A,\alpha_c}^{(i)} = Z_{A',\alpha'_c}^{(i)} = \mathbb{1}$.
2. $Z_{A,\alpha_c}^{(i)} = Z_{A',\alpha'_c}^{(i)} \neq \mathbb{1}$.
3. $Z_{A,\alpha_c}^{(i)} \neq Z_{A',\alpha'_c}^{(i)} \neq \mathbb{1}$.
4. $Z_{A,\alpha_c}^{(i)} \neq Z_{A',\alpha'_c}^{(i)}$ and one of them is an identity.

Let us denote n_{11} as the number of qubits in the 1. case, n_{zz} as the number of qubits in the 2. case, $n_{zz'}$ as the number of qubits in the 3. case, and n_{1z} as the number of qubits in the 4. case. Notice that $n = n_{11} + n_{zz} + n_{zz'} + n_{1z}$. In addition, the relations to the sets of qubits that Z_{A,α_c} and Z_{A',α'_c} act non-trivially on i.e., $\text{sup}(Z_{A,\alpha_c})$ and $\text{sup}(Z_{A',\alpha'_c})$, as well as their associated light cones are as follow:

$$n_{zz} + n_{zz'} = |\text{sup}(Z_{A,\alpha_c}) \cap \text{sup}(Z_{A',\alpha'_c})|, \quad (354)$$

$$n_{1z} = |\text{sup}(Z_{A,\alpha_c}) \cup \text{sup}(Z_{A',\alpha'_c})| - |\text{sup}(Z_{A,\alpha_c}) \cap \text{sup}(Z_{A',\alpha'_c})| \quad (355)$$

$$= \Delta_{Z_A}(\alpha_c) + \Delta_{Z_{A'}}(\alpha'_c) - 2|\text{sup}(Z_{A,\alpha_c}) \cap \text{sup}(Z_{A',\alpha'_c})|. \quad (356)$$

Using these equations from Eq. (352) to Eq. (356), we can compute the covariance term in Eq. (351) as follows

$$\text{Cov}_{\beta} \left[\text{Tr} \left[\rho_{\beta}^{\otimes 2} Z_{A,\alpha_c}^{\otimes 2} \right], \text{Tr} \left[\rho_{\beta}^{\otimes 2} Z_{A',\alpha'_c}^{\otimes 2} \right] \right] = \prod_{i=1}^n \mathbb{E}_{\beta_i} \text{Tr} \left[\rho_{\beta_i}^{\otimes 4} (Z_{A,\alpha_c}^{(i)\otimes 2} \otimes Z_{A',\alpha'_c}^{(i)\otimes 2}) \right] - \prod_{i=1}^n \mathbb{E}_{\beta_i} \text{Tr} \left[\rho_{\beta_i}^{\otimes 2} Z_{A,\alpha_c}^{(i)\otimes 2} \right] \mathbb{E}_{\beta_i} \text{Tr} \left[\rho_{\beta_i}^{\otimes 2} Z_{A',\alpha'_c}^{(i)\otimes 2} \right] \quad (357)$$

$$= \left(\frac{1}{5} \right)^{n_{zz}} \left(\frac{1}{15} \right)^{n_{zz'}} \left(\frac{1}{3} \right)^{n_{1z}} - \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\alpha_c) + \Delta_{Z_{A'}}(\alpha'_c)} \quad (358)$$

$$= \left(\frac{1}{5} \right)^{n_{zz} + n_{zz'}} \left(\frac{1}{3} \right)^{n_{zz'}} \left(\frac{1}{3} \right)^{n_{1z}} - \left(\frac{1}{9} \right)^{n_{zz} + n_{zz'}} \left(\frac{1}{3} \right)^{n_{1z}} \quad (359)$$

$$= \left(\frac{1}{3} \right)^{n_{1z}} \left(\frac{1}{9} \right)^{n_{zz} + n_{zz'}} \left(\left(\frac{9}{5} \right)^{n_{zz} + n_{zz'}} \left(\frac{1}{3} \right)^{n_{zz'}} - 1 \right) \quad (360)$$

$$= \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\alpha_c) + \Delta_{Z_{A'}}(\alpha'_c)} \left[\left(\frac{9}{5} \right)^{|\text{sup}(Z_{A,\alpha_c}) \cap \text{sup}(Z_{A',\alpha'_c})|} \left(\frac{1}{3} \right)^{n_{zz'}} - 1 \right]. \quad (361)$$

Hence, the lower-bound on the purity term can be expressed as

$$\text{Var}_{\theta} [\mathcal{M}(\rho_{\theta}, \rho_{\theta})] \geq \mathbb{E}_{(A,\alpha_c), (A',\alpha'_c) \sim \tilde{\mathcal{D}}_c(\sigma)} \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\alpha_c) + \Delta_{Z_{A'}}(\alpha'_c)} \left[\left(\frac{9}{5} \right)^{|\text{sup}(Z_{A,\alpha_c}) \cap \text{sup}(Z_{A',\alpha'_c})|} \left(\frac{1}{3} \right)^{n(Z_{A,\alpha_c}, Z_{A',\alpha'_c})} - 1 \right] \geq 0, \quad (362)$$

where we introduce a new notation $n(Z_{A,\alpha_c}, Z_{A',\alpha'_c}) := n_{zz'}$ to emphasise the dependence on Z_{A,α_c} and Z_{A',α'_c} . Note that since we begin with $\text{Var}_{\beta} [\mathbb{E}_{\alpha} \mathcal{M}(\rho_{\theta}, \rho_{\theta})]$ in Eq. (342) and we do not do any further lower bound, our final expression is guaranteed to be non-negative. This completes the proof of the supplemental proposition. $\square$

4. Proof of Theorem 3

Proof. We consider the generic form of the MMD variance in Eq. (244) from Supplemental Proposition 3. The lower bound of the MMD variance can be written as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] \geq 4 \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (c_{\sigma}(A) z_A(\tilde{P}))^2 \mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_M} \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\boldsymbol{\alpha}_c)} \quad (363)$$

$$\geq 4 \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (c_{\sigma}(A) z_A(\tilde{P}))^2 \left(\frac{1}{3} \right)^{\mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_M} [\Delta_{Z_A}(\boldsymbol{\alpha}_c)]} \quad (364)$$

$$= 4 \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (c_{\sigma}(A) z_A(\tilde{P}))^2 \left(\frac{1}{3} \right)^{\Delta_{Z_A}^{(\text{avg})}}, \quad (365)$$

where the first inequality is obtained by throwing away the non-negative purity term, and in the second inequality is by applying Jensen's inequality, leading to $\mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_M} \left(\frac{1}{3} \right)^{\Delta_{Z_A}(\boldsymbol{\alpha}_c)} \geq \left(\frac{1}{3} \right)^{\mathbb{E}_{\boldsymbol{\alpha}_c \sim \mathcal{D}_M} [\Delta_{Z_A}(\boldsymbol{\alpha}_c)]}$. The last equality is by using the definition of the average light cone in Eq. (243).

Now, from Proposition 4 the MMD observable acts as few-body for the linear bandwidth $\sigma \in \Theta(n)$ with $c_{\sigma}(A) \in \Omega(1/\text{poly}(n))$ for $|A| \in \mathcal{O}(\log(n))$. Hence, by the assumption that there exists A with $|A| \in \mathcal{O}(\log(n))$ and the average light cone on at most $\log(n)$ qubits together with the polynomial scaling of $z_A^2(\tilde{P})$, that is

$$\Delta_{Z_A}^{(\text{avg})} \in \mathcal{O}(\log(n)) \quad \exists A \text{ such that } |A| \in \mathcal{O}(\log(n)) \text{ and } z_A^2(\tilde{P}) \in \Omega(1/\text{poly}(n)), \quad (366)$$

we have that the MMD variance lower bound scales polynomially with the number of qubits

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] \in \Omega(1/\text{poly}(n)). \quad (367)$$

This completes the proof of the theorem. $\square$

D. Beyond loss gradients - resolving high-order correlations with the MMD

Our results so far (Theorem 2 and Theorem 3) indicate that picking a single bandwidth $\sigma \in \Theta(n)$ maximizes the expected magnitude of gradients for a randomly initialized QCBM. However, while non-vanishing gradients are necessary, they are not sufficient to guarantee reliable training performance.

As discussed in the main text and Supplementary Note IV A, the MMD observable can be decomposed into a weighted sum of Pauli-Z strings ranging from low-body to global interaction terms. For $\sigma \in \Theta(n)$, Proposition 4 ensures that the MMD observable is largely composed of low-body terms, with the contribution from global terms negligible. While this leads to substantial cost gradients, we will argue that losses composed purely of low-body terms struggle to learn global properties of the target distribution. In particular, we will argue that an MMD-type loss that is at most $2k$ bodied cannot distinguish between two distributions with the same marginals on k -qubits but which differ on higher-order marginals. In Supplementary Note IV E we generalise this argument to a broader family of losses for generative modelling.

For a given subset of bits $A \subseteq \mathcal{N} = \{1, 2, \dots, n\}$, denote $\mathbf{x}_A$ as a part of the bitstring $\mathbf{x}$ on that subset A and $\mathbf{x}_{\bar{A}}$ as the rest of the bitstring $\mathbf{x}$. The full bitstring $\mathbf{x}$ can be expressed (not in the right bit order) as $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_{\bar{A}})$. Then, the marginal probability of the training distribution on A can be expressed as

$$\tilde{p}(\mathbf{x}_A) = \sum_{\mathbf{x}_{\bar{A}} \in \{0,1\}^{\otimes(n-|A|)}} \tilde{p}(\mathbf{x}_A, \mathbf{x}_{\bar{A}}) \quad (368)$$

$$= \text{Tr} [\rho_{\tilde{P}} (|\mathbf{x}_A\rangle\langle\mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] , \quad (369)$$

where the sum is over all constellations of the bits that are not in A . In the second line, $\tilde{p}(\mathbf{x}_A)$ is equivalently expressed as the expectation value of a projector onto a computational basis of the subsystem A with the training quantum state $\rho_{\tilde{p}}$. Similarly, the marginal distribution of the model on A is of the form

$$q_{\theta}(\mathbf{x}_A) = \sum_{\mathbf{x}_{\bar{A}} \in \{0,1\}^{\otimes(n-|A|)}} q_{\theta}(\mathbf{x}_A, \mathbf{x}_{\bar{A}}) \quad (370)$$

$$= \text{Tr} [\rho_{\theta} (|\mathbf{x}_A\rangle\langle\mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] . \quad (371)$$

Physically, we note that, when marginals of two distributions agree up to k -bits, this implies that the diagonal elements of the reduced density matrices on any subsets of k qubits are also identical. That is, for all $A \subseteq \mathcal{N}$ such that $|A| \leq k$, if $\tilde{p}(\mathbf{x}_A) = q_{\theta}(\mathbf{x}_A)$, we have

$$\text{Diag}(\text{Tr}_{\bar{A}}[\rho_{\tilde{p}}]) = \text{Diag}(\text{Tr}_{\bar{A}}[\rho_{\theta}]) , \quad (372)$$

where $\bar{A}$ is a complementary of A .

In addition, we recall that the truncated version of the MMD observable defined in Eq. (108) is of the form

$$\tilde{O}_{\text{MMD}}^{(\sigma,k)} = \sum_{l=0}^k \binom{n}{l} (1-p_{\sigma})^{n-l} p_{\sigma}^l D_{2l} \quad (373)$$

$$= \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} (1-p_{\sigma})^{n-|A|} p_{\sigma}^{|A|} \bigotimes_{i \in A} (Z_i \otimes Z_{n+i}) , \quad (374)$$

where in the second line the observable is re-written explicitly as the sum over A (with $l = |A|$). Then, the truncated version of the MMD loss can be expressed as

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\theta) = \text{Tr} [\tilde{O}_{\text{MMD}}^{(\sigma,k)}(\rho_{\theta} \otimes \rho_{\theta})] - 2 \text{Tr} [\tilde{O}_{\text{MMD}}^{(\sigma,k)}(\rho_{\theta} \otimes \rho_{\tilde{p}})] + \text{Tr} [\tilde{O}_{\text{MMD}}^{(\sigma,k)}(\rho_{\tilde{p}} \otimes \rho_{\tilde{p}})] \quad (375)$$

$$= \text{Tr} [\tilde{O}_{\text{MMD}}^{(\sigma,k)}(\rho_{\theta} - \rho_{\tilde{p}})^{\otimes 2}] . \quad (376)$$

We are now ready to state and prove the following proposition.

Proposition 5 (The truncated MMD loss is not faithful). *Consider a distribution $q_{\theta}(\mathbf{x})$ that agrees with the training distribution $\tilde{p}(\mathbf{x})$ on all the marginals up to k bits, but disagrees on higher-order marginals. The distribution $q_{\theta}(\mathbf{x})$ minimizes the truncated MMD loss. That is, suppose*

$$q_{\theta}(\mathbf{x}_A) = \tilde{p}(\mathbf{x}_A) , \quad (377)$$

for all $A \subseteq \{1, 2, \dots, n\}$ with $|A| \leq k$, then

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\theta) = 0 . \quad (378)$$

Crucially, this is true even if for some $B \subseteq \{1, 2, \dots, n\}$ with $|B| > k$

$$q_{\theta}(\mathbf{x}_B) \neq \tilde{p}(\mathbf{x}_B) . \quad (379)$$

Proof. Our proof idea is to express the truncated MMD loss in terms of the marginals up to k bits and show that the loss is minimized when the marginals up to k bits match. First, we explicitly expand $\tilde{O}_{\text{MMD}}^{(\sigma,k)}$ in the truncated MMD loss function in Eq. (376) leading to

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\theta) = \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} (1-p_{\sigma})^{n-|A|} p_{\sigma}^{|A|} [\langle Z_A \rangle_{\theta} - \langle Z_A \rangle_{\tilde{p}}]^2 , \quad (380)$$

where we introduce the shorthand notations $Z_A = \bigotimes_{i \in A} Z_i$, $\langle Z_A \rangle_{\theta} = \text{Tr}[Z_A \rho_{\theta}]$ and $\langle Z_A \rangle_{\tilde{p}} = \text{Tr}[Z_A \rho_{\tilde{p}}]$. By expressing Z_A in the computational basis, we have

$$Z_A = \sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} |\mathbf{x}_A\rangle \langle \mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}}. \quad (381)$$

So, the expectation of Z_A can be written as a sum of the marginals probabilities on A with the definition in Eq. (368) as follows,

$$\langle Z_A \rangle_{\tilde{p}} = \sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} \tilde{p}(\mathbf{x}_A), \quad (382)$$

and

$$\langle Z_A \rangle_{\theta} = \sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} q_{\theta}(\mathbf{x}_A). \quad (383)$$

Then, we have

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma, k)}(\theta) = \sum_{\substack{A \in \mathcal{N} \\ |A| \leq k}} (1 - p_{\sigma})^{n-|A|} p_{\sigma}^{|A|} \left[\sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} (\tilde{p}(\mathbf{x}_A) - q_{\theta}(\mathbf{x}_A)) \right]^2 \quad (384)$$

$$= 0, \quad (385)$$

which completes the proof. Note that we do not need information of the marginals beyond k bits and therefore this leads to the unfaithfulness in the sense that higher-order marginals can disagree even with the truncated loss being minimized. $\square$

In Proposition 5, we show that if the marginals between the model and training distributions match up to k bits, then the truncated loss of order k is minimized with the model distribution. We now show that the inverse direction also holds. That is, minimizing the truncated loss means learning the marginals of the training distribution.

To show this, we consider again the truncated MMD loss in Eq (380) and notice that the loss is minimized and equals to 0 if and only if

$$\langle Z_A \rangle_{\theta} = \langle Z_A \rangle_{\tilde{p}}, \quad (386)$$

for all $A \in \mathcal{N}$ such that $|A| \leq k$. Concerning the marginal probabilities, we now decompose the projector $|\mathbf{x}_A\rangle \langle \mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}}$ in the Pauli basis

$$|\mathbf{x}_A\rangle \langle \mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}} = \bigotimes_{i \in A} \frac{1}{2} (\mathbb{1}_i + (-1)^{x_i} Z_i) \quad (387)$$

$$= \frac{1}{2^{|A|}} \sum_{B \subseteq A} (-1)^{\sum_{i \in B} x_i} Z_B, \quad (388)$$

where we expanded the product fully with $Z_B = \bigotimes_{i \in B} Z_i$, where B are all possible subsets of A . Thus any k bit marginal can be computed from a sum of the average parities of all subsets up to k bits via

$$\tilde{p}(\mathbf{x}_A) = \frac{1}{2^{|A|}} \sum_{B \subseteq A} (-1)^{\sum_{i \in B} x_i} \langle Z_B \rangle_{\tilde{p}}. \quad (389)$$

It is clear from Eq. (380) that training on the k -truncated MMD learns all average parities of the target distribution up to k bits, Eq. (386), and hence Eq. (389) implies we also learn all marginals up to and including k bits. Put another way, the difference between model and training marginal probabilities is given by

$$\tilde{p}(\mathbf{x}_A) - q_{\theta}(\mathbf{x}_A) = \frac{1}{2^{|A|}} \sum_{B \subseteq A} (-1)^{\sum_{i \in B} x_i} (\langle Z_B \rangle_{\tilde{p}} - \langle Z_B \rangle_{\theta}), \quad (390)$$

$$= 0, \quad (391)$$

for all A such that $|A| \leq k$.

E. Distinguishing marginals using an arbitrary loss

In Supplementary Note IV D, we have shown that a truncated MMD loss operator cannot distinguish between model distributions that agree with the data distribution up until a certain order of marginals, but disagree beyond. In this section, we show that this phenomenon can be extended to general generative losses for classical data that can be formulated as the expectation value of some observable. As a key example, we first consider loss functions $\mathcal{L}(\theta) = \text{Tr}[O\rho_\theta]$ with an observable in the following form,

$$O = \sum_{\mathbf{x} \in \mathcal{X}} D_\alpha(\mathbf{x}) |\mathbf{x}\rangle\langle\mathbf{x}|. \quad (392)$$

Here, $D_\alpha(\mathbf{x})$ is the eigenvalue of the operator corresponding with the computational basis sample $\mathbf{x}$, which could additionally be parametrized by α . Notably, the loss for the Generator in a quantum GAN can be expressed in this form, where $D_\alpha(\mathbf{x})$ is the classification output of the Discriminator.

The truncated version of O in the Pauli basis up to k -body terms can be expressed as

$$O^{(k)} = \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} c_A Z_A, \quad (393)$$

where $c_A = \frac{1}{2^n} \sum_{\mathbf{x}} D_\alpha(\mathbf{x}) (-1)^{\sum_{i \in A} x_i}$, and $Z_A = \bigotimes_{i \in A} Z_i$ are Pauli operators acting non-trivially on qubits i in a subset of qubits A .

Now, we show that the loss assigned by the truncated loss function is the same between that the model state ρ_{θ_1} , which matches the training distribution exactly (i.e., it is the global optimum of the full loss but not necessarily of the truncated one), and a state ρ_{θ_2} , matches the training distribution up to k -bit marginals but disagrees beyond. Both are characterized by the property

$$\text{Tr}[\rho_{\theta_1} (|\mathbf{x}_A\rangle\langle\mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] = \text{Tr}[\rho_{\theta_2} (|\mathbf{x}_A\rangle\langle\mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] = \tilde{p}(\mathbf{x}_A). \quad (394)$$

for all $A \subseteq \mathcal{N}$ such that $|A| \leq k$. This implies that the reduced states $\rho_{\theta_1, A} = \text{Tr}_{\bar{A}}[\rho_{\theta_1}]$, $\rho_{\theta_2, A} = \text{Tr}_{\bar{A}}[\rho_{\theta_2}]$ and $\rho_{\tilde{p}, A} = \text{Tr}_{\bar{A}}[\rho_{\tilde{p}}]$ have the same diagonal, that is $\text{Diag}(\rho_{\theta_1, A}) = \text{Diag}(\rho_{\theta_2, A}) = \text{Diag}(\rho_{\tilde{p}, A})$. Consequently, the expectations of any diagonal Pauli operator Z_A on A are the same,

$$\langle Z_A \rangle_{\theta_1} = \langle Z_A \rangle_{\theta_2} = \langle Z_A \rangle_{\tilde{p}}. \quad (395)$$

Now consider the difference in the loss between any of these two states i.e., $\rho, \rho' \in \{\rho_{\theta_1}, \rho_{\theta_2}, \rho_{\tilde{p}}\}$

$$d(\rho, \rho') = \mathcal{L}(\rho) - \mathcal{L}(\rho') \quad (396)$$

$$= \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} c_A (\text{Tr}_A[Z_A \rho] - \text{Tr}_A[Z_A \rho']) \quad (397)$$

$$= 0. \quad (398)$$

This shows that any loss composed exclusively of low-body terms cannot distinguish between distributions with the same low-order marginals (but potentially different higher-order marginals). Beyond losses that are natively composed entirely of low-body operators, this result becomes of practical importance when the loss $\mathcal{L}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} \text{Tr}[D_\alpha(\mathbf{x}) |\mathbf{x}\rangle\langle\mathbf{x}| \rho_\theta]$ is effectively composed entirely of low-body terms, i.e., if the global contributions to $\mathcal{L}(\theta)$ are too small to be resolved using the available shot budget. In that case, $\mathcal{L}(\theta)$ is well approximated by its truncated version and our argument applies. Whether this is or is not the case is determined by the choice in $D_\alpha(\mathbf{x})$.

We note that in this derivation we left the structure of D_α entirely general up to the constraint that it is diagonal in the computational basis and acts only on a single copy of the model distribution at a time. But our proof can be directly applied to general generative losses $\mathcal{L}_{gen}(\theta, \{C_O\})$ for classical data which take expectation values of several observables $C_O = \text{Tr}[O \rho_\theta^{\otimes m}]$ with operators O acting on m different sub-systems, each of which contains up to k -body terms in the Pauli basis. That is, we have the general operator of the form

$$O = \sum_{\substack{A_1, \dots, A_m \subseteq \mathcal{N} \\ |A_1|, \dots, |A_m| \leq k}} c_{A_1, \dots, A_m}(\alpha, \tilde{P}) (Z_{A_1} \otimes Z_{A_2} \otimes \dots \otimes Z_{A_m}), \quad (399)$$

where, $A_1, A_2, \dots, A_m$ are subsets of $\mathcal{N} = \{1, \dots, n\}$, $c_{A_1, \dots, A_m}(\boldsymbol{\alpha}, \tilde{P})$ are real coefficients that can depend on training data, and $Z_{A_j} = \bigotimes_{i \in A_j} Z_{(j-1)n+i}$ acting non-trivially on the qubits of the j^{th} subsystem.

This form of the general loss covers loss functions for quantum circuit Born machines (in particular the MMD, as outlined in Sec. “Implicit losses: maximum mean discrepancy” in the main text), quantum GANS, quantum Boltzmann machines [11], and any other proposed (quantum) generative model on classical discrete data. Therefore, any diagonal loss operator that implements a generative modelling loss and contains only low-bodied operators cannot be used to reliably learn global probability marginals.

V. SUPPLEMENTARY NOTE - ANALYSIS ON THE QUANTUM FIDELITY LOSS

In this section, we present an approach that can be used to estimate the local fidelity quantity $\mathcal{L}_{QF}^{(L)}(\boldsymbol{\theta})$ using a series of Hadamard tests without explicitly loading the training data into a quantum state or requiring a quantum oracle.

We begin by introducing a pure quantum state corresponding to the training dataset $|\phi\rangle = \sum_{\mathbf{x}} \sqrt{\tilde{p}(\mathbf{x})} |\mathbf{x}\rangle$. Learning the training distribution is equivalent to the state learning task with $|\phi\rangle$ as the target state. We note that our choice of having the coefficient as $\sqrt{\tilde{p}(\mathbf{x})}$ is arbitrary and, generally, any target quantum state with the probabilities corresponding to the training probabilities would be valid candidates for the task.

As discussed in the main text, the quantum fidelity can be used as a cost function in this learning task

$$\mathcal{L}_{QF}(\boldsymbol{\theta}) = 1 - |\langle \phi | \psi(\boldsymbol{\theta}) \rangle|^2. \quad (400)$$

However, the globality of the loss leads to barren plateaus, which in turn leads to the untrainability of the loss. However, the local version of the quantum fidelity has been shown to be both trainable with the shallow depth circuits and faithful to the original global version [12]. Specifically, the local quantum fidelity is of the form

$$\mathcal{L}_{QF}^{(L)}(\boldsymbol{\theta}) = 1 - \langle \phi | U(\boldsymbol{\theta}) H_L U^\dagger(\boldsymbol{\theta}) | \phi \rangle, \quad (401)$$

with

$$H_L = \frac{1}{n} \sum_{i=1}^n |0_i\rangle\langle 0_i| \otimes \mathbb{1}_{\bar{i}}, \quad (402)$$

where $|0_i\rangle\langle 0_i|$ are now single-qubit projectors. In the context of the generative modeling with classical, there is an additional challenge as only the training dataset $\tilde{P}$ is given to us and not the state $|\phi\rangle$. Using Hadamard tests, we now show that measuring the local quantum fidelity can be achieved efficiently without loading the training data into the quantum state. To see this, we first express H_L in the Pauli basis

$$H_L = \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathbb{1}_i + Z_i}{2} \right) \otimes \mathbb{1}_{\bar{i}} \quad (403)$$

$$= \frac{\mathbb{1}}{2} + \frac{1}{2n} \sum_i Z_i. \quad (404)$$

We then expand $\mathcal{L}_{QF}^{(L)}(\boldsymbol{\theta})$ as

$$\mathcal{L}_{QF}^{(L)}(\boldsymbol{\theta}) = 1 - \langle \phi | U(\boldsymbol{\theta}) \left(\frac{\mathbb{1}}{2} + \frac{1}{2n} \sum_i Z_i \right) U^\dagger(\boldsymbol{\theta}) | \phi \rangle \quad (405)$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_i \langle \phi | U(\boldsymbol{\theta}) Z_i U^\dagger(\boldsymbol{\theta}) | \phi \rangle \quad (406)$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_{i, \mathbf{x}, \mathbf{x}'} \sqrt{\tilde{p}(\mathbf{x}) \tilde{p}(\mathbf{x}')} \langle \mathbf{x} | U(\boldsymbol{\theta}) Z_i U^\dagger(\boldsymbol{\theta}) | \mathbf{x}' \rangle \quad (407)$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_{i, \mathbf{x}, \mathbf{x}'} \sqrt{\tilde{p}(\mathbf{x}) \tilde{p}(\mathbf{x}')} \langle \mathbf{x} | U(\boldsymbol{\theta}) Z_i U^\dagger(\boldsymbol{\theta}) U_{\mathbf{x}', \mathbf{x}} | \mathbf{x} \rangle \quad (408)$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_{i, \mathbf{x}, \mathbf{x}'} \sqrt{\tilde{p}(\mathbf{x}) \tilde{p}(\mathbf{x}')} \langle \mathbf{x} | \tilde{U}(\boldsymbol{\theta}, i, \mathbf{x}, \mathbf{x}') | \mathbf{x} \rangle, \quad (409)$$

where in the third line we explicitly expand $|\phi\rangle = \sum_{\mathbf{x}} \sqrt{p(\mathbf{x})} |\mathbf{x}\rangle$, and in the fourth line we introduce $U_{\mathbf{x}', \mathbf{x}}$ which is the unitary mapping from the computational basis $\mathbf{x}$ to $\mathbf{x}'$ by applying the necessary single-qubit flips. Finally, in the last line, we introduce $\tilde{U}(\boldsymbol{\theta}, i, \mathbf{x}, \mathbf{x}') = U(\boldsymbol{\theta}) Z_i U^\dagger(\boldsymbol{\theta}) U_{\mathbf{x}', \mathbf{x}}$ which summarizes the full unitary to be implemented for any pair $\mathbf{x}$ and $\mathbf{x}'$. Each term can now be estimated using two Hadamard tests, i.e., one for the real part and the one for the imaginary part of each overlap. If $N_p \in \mathcal{O}(\text{poly}(n))$ is the number of unique bitstrings in the training dataset $\tilde{P}$, one thus requires $2nN_p^2 \in \mathcal{O}(\text{poly}(n))$ Hadamard tests to evaluate all terms in the loss. The number of Hadamard tests can be reduced by a factor of 2 if the quantum model is constructed to span only either real or imaginary subspace. The number of controlled unitaries can also be reduced to simple control phase gates by using a diagonal ansatz [13]. While the number of Hadamard tests naively scales with the number of training bitstrings, this overhead is expected to be significantly reduced by employing techniques such as stochastic gradient descent [14] which allows us to stochastically optimize the loss in an unbiased manner.

VI. SUPPLEMENTARY NOTE - ADDITIONAL TRAINING WITH EXPONENTIAL SUPPORT

As stated in the main text, QCBMs are not expected to be able to learn distribution with exponential support when the training data are stored in classical computers. This is since even storing so much data is unrealistic beyond a few dozens of qubits. However, due to the focus on systems with very few qubits, most applications taken from the literature utilize on distribution with non-zero probabilities on a macroscopic number of bitstrings. Such cases can still be trainable with the KL divergence and finite shots if the number of qubits is small enough ($n \leq 12$). We perform the same training procedure as in the main text in Sec. “Training on a HEP dataset” in the main text, with the additions of gradient batching over $k = 10$ iterations, and gradient clipping with a threshold of $\tau = 0.1$. These details aim at stabilising the optimisation and improve the performance, and follow best practices. Fig. 2 shows this numerically on the ECAL dataset with $n = 6$ qubits. More particularly, we train on a transformed version of the dataset that follows $-\log(\tilde{p}(\mathbf{x}))$, which exhibits exponential support (top three rows) and additionally on the original probabilities (bottom three rows), which have only a polynomial support, for different number of shots (10^2 , 10^3 , 10^4 and ∞) and layers (1, 8 and 16). Each column is divided into two blocks of three plots each, where the first shows the TVD during the training, the second the generated histograms with the best ansatz against the target distribution, and the third the absolute error between the two distributions. We observe that the three loss functions are comparable for the exponential support case, while KLD breaks down when the support is polynomial and using fewer than 10^3 shots.

We recall that these numerics are not a contradiction with the message of this paper, since learning a distribution with exponential support is not scalable using QCBMs, but may explain why the exponential concentration issue has

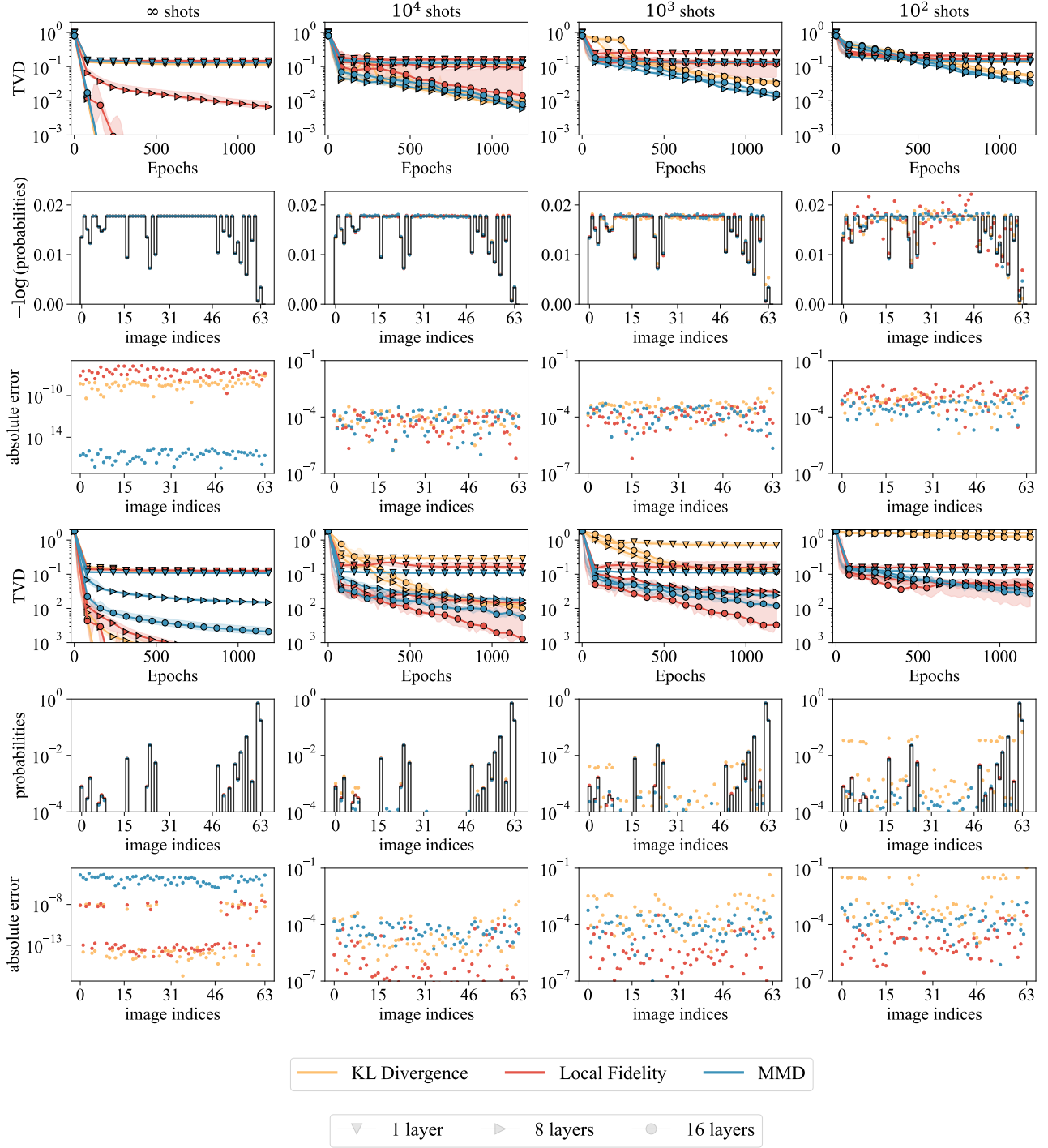


Figure 2. Training on the ECAL dataset with $n = 6$ qubits on a dataset with exponential support (top three rows) and one with polynomial support (last three rows). The first rows shows the TVD during training, while the second displays the generated distribution against the target one (black) and the third the absolute error between the two.

not been discussed for quantum generative models before.

-
- [1] Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti, “An initialization strategy for addressing barren plateaus in parametrized quantum circuits,” *Quantum* **3**, 214 (2019).
 - [2] Sergio Boixo, Sergei V Isakov, Vadim N Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J Bremner, John M Martinis, and Hartmut Neven, “Characterizing quantum supremacy in near-term devices,” *Nature Physics* **14**, 595–600 (2018).
 - [3] Milton Abramowitz, “Abramowitz and stegun: Handbook of mathematical functions,” US Department of Commerce **10** (1972).
 - [4] Sean Mullane, “Sampling random quantum circuits: a pedestrian’s guide,” *arXiv preprint arXiv:2007.07872* (2020).
 - [5] Hubert de Guise, Olivia Di Matteo, and Luis L Sánchez-Soto, “Simple factorization of unitary transformations,” *Physical Review A* **97**, 022328 (2018).
 - [6] Zoë Holmes, Andrew Arrasmith, Bin Yan, Patrick J. Coles, Andreas Albrecht, and Andrew T Sornborger, “Barren plateaus preclude learning scramblers,” *Physical Review Letters* **126**, 190501 (2021).
 - [7] Daniel A Roberts and Beni Yoshida, “Chaos and complexity by design,” *Journal of High Energy Physics* **2017**, 121 (2017).
 - [8] Matthias C Caro, Hsin-Yuan Huang, Nicholas Ezzell, Joe Gibbs, Andrew T Sornborger, Lukasz Cincio, Patrick J Coles, and Zoë Holmes, “Out-of-distribution generalization for learning quantum dynamics,” *Nature Communications* **14**, 3751 (2023).
 - [9] Alistair Letcher, Stefan Woerner, and Christa Zoufal, “Tight and efficient gradient bounds for parameterized quantum circuits,” *arXiv preprint arXiv:2309.12681* (2023).
 - [10] Antonio Anna Mele, “Introduction to haar measure tools in quantum information: A beginner’s tutorial,” *Quantum* **8**, 1340 (2024).
 - [11] Mohammad H. Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko, “Quantum boltzmann machine,” *Phys. Rev. X* **8**, 021050 (2018).
 - [12] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles, “Quantum-assisted quantum compiling,” *Quantum* **3**, 140 (2019).
 - [13] Benjamin Commeau, M. Cerezo, Zoë Holmes, Lukasz Cincio, Patrick J. Coles, and Andrew Sornborger, “Variational Hamiltonian diagonalization for dynamical quantum simulation,” *arXiv preprint arXiv:2009.02559* (2020).
 - [14] Ryan Sweke, Frederik Wilde, Johannes Jakob Meyer, Maria Schuld, Paul K Fährmann, Barthélémy Meynard-Piganeau, and Jens Eisert, “Stochastic gradient descent for hybrid quantum-classical optimization,” *Quantum* **4**, 314 (2020).