

# Trainability barriers and opportunities in quantum generative modeling

Manuel S. Rudolph,<sup>1,\*</sup> Sacha Lerch,<sup>1,\*</sup> Supanut Thanasilp,<sup>1,2,\*</sup>  
 Oriël Kiss,<sup>3,4</sup> Sofia Vallecorsa,<sup>3</sup> Michele Grossi,<sup>3</sup> and Zoë Holmes<sup>1</sup>

<sup>1</sup>*Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

<sup>2</sup>*Chula Intelligent and Complex Systems, Department of Physics,*

*Faculty of Science, Chulalongkorn University, Bangkok, Thailand, 10330*

<sup>3</sup>*European Organization for Nuclear Research (CERN), Geneva 1211, Switzerland*

<sup>4</sup>*Department of Nuclear and Particle Physics, University of Geneva, Geneva 1211, Switzerland*

(Dated: May 5, 2023)

Quantum generative models, in providing inherently efficient sampling strategies, show promise for achieving a near-term advantage on quantum hardware. Nonetheless, important questions remain regarding their scalability. In this work, we investigate the barriers to the trainability of quantum generative models posed by barren plateaus and exponential loss concentration. We explore the interplay between explicit and implicit models and losses, and show that using implicit generative models (such as quantum circuit-based models) with explicit losses (such as the KL divergence) leads to a new flavour of barren plateau. In contrast, the Maximum Mean Discrepancy (MMD), which is a popular example of an implicit loss, can be viewed as the expectation value of an observable that is either low-bodied and trainable, or global and untrainable depending on the choice of kernel. However, in parallel, we highlight that the low-bodied losses required for trainability cannot in general distinguish high-order correlations, leading to a fundamental tension between exponential concentration and the emergence of spurious minima. We further propose a new local quantum fidelity-type loss which, by leveraging quantum circuits to estimate the quality of the encoded distribution, is both faithful and enjoys trainability guarantees. Finally, we compare the performance of different loss functions for modelling real-world data from the High-Energy-Physics domain and confirm the trends predicted by our theoretical results.

## I. INTRODUCTION

The advent of quantum computing has opened up new avenues for solving classically intractable problems [1–4]. Naturally, researchers gravitate towards finding the first high-value applications that could be tackled with near- and mid-term quantum devices [5]. This includes not only speed-ups [3, 6–8], but potentially superior memory efficiency [9] or concrete qualitative improvements [10, 11]. Quantum machine learning (QML) is one of the domains that attracts this attention [2]. Quantum systems, in being inherently probabilistic, are particularly well suited to generative modelling tasks [12]. Generative models aim to learn the underlying distribution of a dataset and thereby provide a means of generating new data samples that are similar to the original data. As well as providing a naturally efficient means of generating samples, quantum generative models can provably encode probability distributions that are out of reach for classical models [13–15], and have been proposed for various applications, such as handwritten digits [16], finance [17] or High-Energy-Physics [18].

Despite the excitement surrounding the potential of generative QML, there remain substantial questions concerning its scalability. This is non-trivial to assess since implementations are constrained by hardware limitations to small-scale proof-of-principle problems [16, 17, 19–21]. Thus analytic results are essential to guide the successful development of this field. Of particular concern is the growing body of literature on cost function concentration and barren plateaus [22–29], where loss function values can exponentially concentrate around a fixed value and loss gradients vanish exponentially with growing problem size. This phenomenon, which exponentially increases the resources required for training, originates from different sources [22, 24, 30–38], and has been studied in a number of architectures [22, 24, 29, 39–44] as well as classes of cost function [31, 36, 39]. However, its impact on quantum generative modelling has thus far been largely overlooked.

In this work, we provide a thorough study of trainability barriers and opportunities in quantum generative modelling. Critical to our analysis is the distinction between explicit and implicit models and losses. Explicit models provide efficient access directly to the model probabilities, whereas implicit models only provide samples drawn from their distribution [45]. Quantum circuit Born machines (QCBMs) [46], the focus of this work, encode a probabil-

---

\* The first three authors contributed equally to this work.

Circuit depth	Explicit loss (pairwise)		Implicit loss (MMD)
	Conventional strategy	Quantum strategy	
Product	No (Corollary 2)	Yes (Local Quantum Fidelity [31])	Yes ( $\sigma \in \Theta(n)$ , Theorem 2)
Shallow			Yes ( $\sigma \in \Theta(n)$ , Conjecture 1)
Deep		No [22, 30]	No [22, 30]

Table I. **Summary of our main results.** This table summarizes our key analytical results on the trainability of different loss functions in quantum generative modelling tasks. Without a strong inductive bias, pairwise explicit losses are untrainable for all circuit depths with the conventional sampling strategy. A quantum strategy could be utilised to efficiently estimate the local quantum fidelity, Eq. (48), which is trainable for a shallow-depth circuit. The MMD using a classical Gaussian kernel with a linearly-scaled bandwidth ( $\sigma \in \Theta(n)$ ) is expected to be trainable for a shallow-depth circuits. Note that ‘Yes’ here indicates the existence of regimes with trainability guarantees- it does not preclude untrainable regimes including, for example, the use of global quantum fidelity or the MMD with a fixed bandwidth.

ity distribution in an  $n$ -qubit pure state and thus are a paradigmatic example of an implicit model. Mirroring the capabilities of the models, explicit losses are those that are formulated explicitly in terms of the model and target probabilities, whereas implicit losses compare samples from the model and the training distribution. The most commonly used explicit loss for quantum generative models is the Kullback-Leibler (KL) divergence [47]. Other examples include the Jensen-Shannon divergence (JSD), the total variation distance (TVD) and the classical fidelity. The Maximum Mean Discrepancy (MMD) [48] on the other hand is one of the leading examples of an implicit loss.

Here we argue that the tension between using an implicit generative model (providing only samples) with an explicit loss (requiring access to probabilities) leads to a new flavour of barren plateau. This result disqualifies all before-mentioned explicit losses, and crucially the KL divergence, for efficient training of QCBMs without a strong inductive bias towards the target distribution. In contrast, the MMD as an implicit loss exhibits more nuanced behaviour and can be either trainable or untrainable. By viewing the classical MMD loss as the expectation value of a quantum observable, we show that varying the bandwidth parameter of a Gaussian kernel interpolates the MMD loss between a loss composed of predominantly global terms

and one composed of low-bodied terms with either exponentially or polynomially decaying loss variances in the number of qubits. These results are summarised in Table I.

In parallel, we provide insights into how the globality of a generative loss affects the types of correlations in a dataset that can reliably be learned. In particular, we show that a  $k$ -bodied loss (see Fig. 4) cannot distinguish between distributions that agree on all  $k$ -marginals but disagree about higher-order correlations. Hence we argue that in the context of quantum generative modelling it is advantageous to train on *full-bodied* losses, that is losses containing both low and high-bodied terms, rather than the purely local losses advocated elsewhere in quantum machine learning. The MMD is then a promising candidate choice for the training of QCBMs as its bodyness can be controlled via the bandwidth parameter.

We additionally expand the pool of viable loss functions by proposing a new local quantum fidelity-type loss which leverages what we call a quantum strategy for evaluating losses. This is to be contrasted with the conventional measurement strategy which simply uses samples from the model distribution in the computational basis. We provide an efficient training protocol using the local quantum fidelity loss with provable trainability guarantees.

Finally, we support our analysis with a comparison of the performance of the KL divergence, MMD and local quantum fidelity losses for modelling High-Energy-Physics (HEP) data. Specifically, we consider electron energy depositions in the electromagnetic calorimeter (ECAL) part of detectors involved in a typical proton-proton collision experiment at the LHC. We learn to generate hits in the detector as black and white images of various sizes, with up to 16 qubits. We confirm that the properly-tuned MMD and the local quantum fidelity losses remain trainable using a restrictive shot budget, while training with the KL divergence becomes increasingly futile.

## II. FRAMEWORK

The goal of generative modelling is to use samples from a target distribution  $p(\mathbf{x})$  to learn a model of  $p(\mathbf{x})$  which can be used to generate new samples. More concretely, as sketched in Fig. 1, a generative model takes as input a training dataset  $\tilde{P}$  consisting of  $M = |\tilde{P}|$  samples drawn from the target distribution  $p(\mathbf{x})$ . This training set can be used to construct the empirical probability distribution  $\tilde{p}(\mathbf{x})$  for all samples  $\mathbf{x} \in \tilde{P}$ . The training dataset, or the training distribution, is then used to train the variational parameters  $\theta$  of a parameterized probability distribution  $q_{\theta}(\mathbf{x})$ . If successful, the output of the algorithm is a set

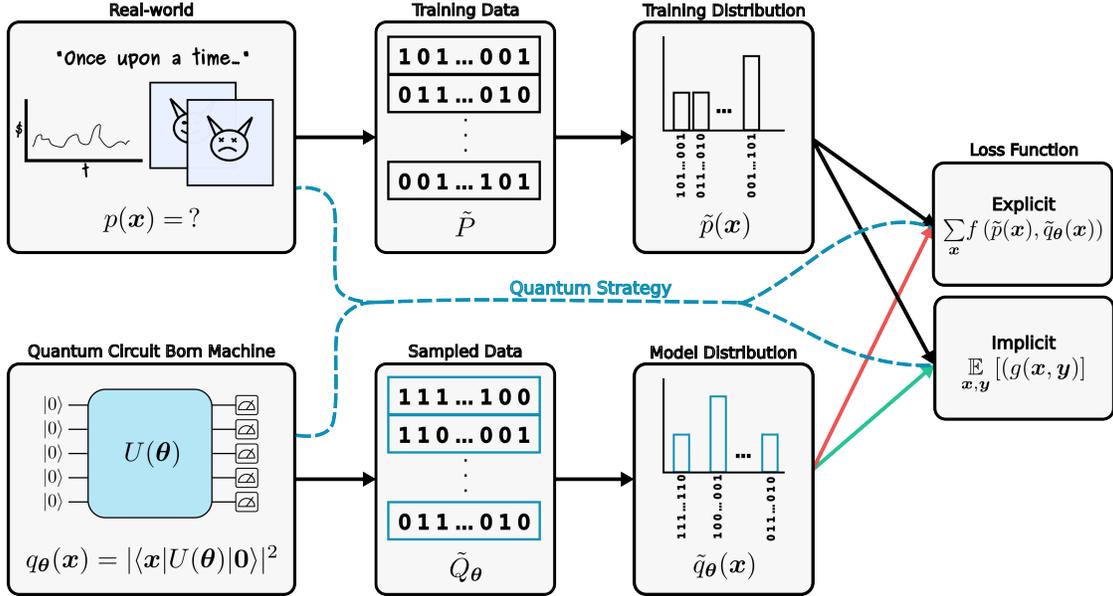


Figure 1. **The generative modelling framework using quantum circuit Born machines.** Given a training dataset  $\tilde{P}$  with distribution  $\tilde{p}(\mathbf{x})$  over discrete data samples  $\mathbf{x}$ , the goal of a QCBM is to learn a distribution  $q_{\theta}(\mathbf{x})$  which models the real-world distribution  $p(\mathbf{x})$  from which the training data itself was sampled. This is done by tuning the parameters  $\theta$  of a parametrized quantum circuit such that the QCBM minimizes a loss function that estimates the distance between the model and the training distribution. The QCBM is an implicit model and can thus in general not be paired with an explicit loss function, but it may be trainable using an implicit loss. In contrast to the conventional loss estimation strategy (solid lines) of generating a set of samples  $\tilde{Q}_{\theta}$  and forming an empirical distribution  $\tilde{q}_{\theta}(\mathbf{x})$ , strategies that are ‘more quantum’ (dashed lines) can be employed with the aim of allowing QCBMs to be trained with loss functions which conventionally appear explicit.

of optimized parameters  $\theta_{\text{opt}}$  such that the trained model  $q_{\theta_{\text{opt}}}(\mathbf{x})$  well-approximates the unknown target distribution  $p(\mathbf{x})$ . The trained model  $q_{\theta_{\text{opt}}}(\mathbf{x})$  can then be used to generate new and previously unseen data. For compactness, we use the notation  $p$  and  $q_{\theta}$  to denote the target and model distributions respectively.

The process of training requires a *loss function*  $\mathcal{L}(\theta)$  which estimates the distance between the model distribution  $q_{\theta}(\mathbf{x})$  and the training distribution  $\tilde{p}(\mathbf{x})$ . For typical choices in loss function (detailed further in Section II B), the loss is minimised when the model parameters  $\theta$  are tuned such that the model distribution perfectly matches the empirical distribution obtained from the training data. That is,  $\mathcal{L}(\theta) = 0$  if and only if  $q_{\theta}(\mathbf{x}) = \tilde{p}(\mathbf{x})$  over the entire data space  $\mathcal{X}$ . Thus, by perfectly minimizing the loss, one perfectly learns the empirical distribution  $\tilde{p}(\mathbf{x})$  but not the true target distribution  $p(\mathbf{x})$ . This scenario is commonly called *overfitting*<sup>1</sup>. To allow for *generaliza-*

*tion* [49], whereby the model can generate novel data with similar properties to the training data, one seeks to significantly reduce (but not perfectly minimize) the training loss. While generalization is the end-all goal of generative models, it is not the focus of this work. Instead, we focus on the training component of the generative framework, as failing to train also prohibits generalization.

### A. Quantum circuit models

One prototypical quantum generative model is the *quantum circuit Born machine* (QCBM) [13, 46, 50, 51]. Owing its name to the Born rule of quantum mechanics, a QCBM encodes a probability distribution over discrete data (here bitstrings) in an  $n$ -qubit pure quantum state that depends on a parameterized unitary  $U(\theta)$ ,

$$q_{\theta}(\mathbf{x}) = |\langle \mathbf{x} | U(\theta) | 0 \rangle|^2. \quad (1)$$

Here  $|\mathbf{x}\rangle$  is a computational basis state corresponding to a bitstring  $\mathbf{x}$  and, without loss of generality, an initial state can be chosen as  $|0\rangle = |0\rangle^{\otimes n}$ . We note that estimating

<sup>1</sup> In contrast, discriminative machine learning models can be perfectly minimized on the training data and not be overfitted.

$q_{\theta}(\mathbf{x})$  is equivalent to finding the expectation value of a global projector  $|\mathbf{x}\rangle\langle\mathbf{x}|$ . More fundamentally, QCBMs enable the encoded distribution to be efficiently sampled simply by measuring in a chosen computational basis. That is, every measurement of the quantum state provides an unbiased sample from the encoded distribution (in an ideal noise-free setting). This is a very desirable property in generative models that many (classical) generative models do not share with the QCBM. Sampling techniques for classical generative models are often unreliable and may break down for certain distributions, as is the case for *restricted Boltzmann machines* (RBMs) [52, 53]. Born machines represent an effort to create a powerful, flexible and efficient generative model for classical discrete data, and as well as numerous ‘standard’ digital quantum implementations [16, 17, 19–21], they have been widely implemented using tensor networks [54–57], continuous variable hardware [58], in a conditional setting [59, 60], with non-linearities [61].

An important, but rather subtle, distinction in generative modelling is that between *explicit and implicit generative models* [45, 62]. Explicit generative models are ones that allow efficient access to the model probability  $q_{\theta}(\mathbf{x})$  for any data sample  $\mathbf{x}$ . Here, ‘efficient’ means that the probabilities can be computed in a time and memory that are polynomial in the size of the data samples, i.e.,  $\mathcal{O}(\text{poly}(n))$  resources. Explicit (classical) generative models include for example auto-regressive models [63], RNNs [64], tensor networks without loops (which includes tensor network Born machines) [54, 55], and many forms of density estimators. In contrast, *implicit* models lack this property and instead offer efficient access to samples from  $q_{\theta}(\mathbf{x})$ , which some forms of explicit models may struggle with. A popular example of an implicit generative model are *Generative Adversarial Networks* (GANs) [65] that leverage an implicit training scheme to learn powerful generators.

In the case of QCBMs implemented on quantum devices, it becomes evident that we do not have (efficient) explicit access to  $q_{\theta}(\mathbf{x})$ , but only to samples of the distribution in the computational basis. Consequently, QCBMs can be classified as implicit generative models. In this work, we study the trainability issues that QCBMs suffer from as a result.

## B. Loss functions

Similarly to the distinction between explicit and implicit generative models, we draw a distinction between *explicit and implicit loss functions*. In broad terms, explicit losses are those that can only be formulated explicitly in terms of

the target and model *probabilities*, whereas implicit losses are those that can be formulated in terms of an average over model and training data *samples*. This distinction at the level of loss functions thus mirrors the capabilities and limitations of explicit and implicit generative models.

More concretely, we define an *explicit loss* as a loss function  $\mathcal{L}$  that can be written solely as a function of the probabilities of the target and model distributions, without any dependence on the data itself. Explicit losses thus take the general form

$$\mathcal{L}_{\text{expl}}(\theta) := \sum_{\mathbf{x}_1 \dots \mathbf{x}_r} f\left(p(\mathbf{x}_1), \dots, p(\mathbf{x}_r), q_{\theta}(\mathbf{x}_1), \dots, q_{\theta}(\mathbf{x}_r)\right), \quad (2)$$

where  $f(\cdot)$  is a function that depends on the target probabilities  $p(\mathbf{x}_i)$  and model probabilities  $q_{\theta}(\mathbf{x}_i)$  for data variables  $\mathbf{x}_i \in \mathcal{X}$  with  $i = 1, \dots, r$ . For this loss to be useful, the function  $f$  should be chosen such that it measures the distance between the probability distributions  $p$  and  $q_{\theta}$ . Crucially, the function  $f$  does not take the data values  $\mathbf{x}$  themselves as arguments.

While in full generality explicit losses could compare multiple copies of the target and model probabilities (i.e., we can have  $r > 1$ ), in practice, they usually take the simpler form

$$\mathcal{L}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} f(p(\mathbf{x}), q_{\theta}(\mathbf{x})). \quad (3)$$

We call such losses *pairwise explicit losses* since they compare the model and target probabilities on the same data samples, or in our case, bitstrings. The pairwise explicit loss covers all so-called  $f$ -divergences [66], including the commonly encountered KL divergence (KLD) [67],

$$\mathcal{L}^{\text{KLD}}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log \left( \frac{p(\mathbf{x})}{q_{\theta}(\mathbf{x})} \right), \quad (4)$$

the reverse-KLD,

$$\mathcal{L}^{\text{rev-KLD}}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} q_{\theta}(\mathbf{x}) \log \left( \frac{q_{\theta}(\mathbf{x})}{p(\mathbf{x})} \right), \quad (5)$$

the Jensen-Shannon divergence (JSD) [68],

$$\mathcal{L}^{\text{JSD}}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} \left[ p(\mathbf{x}) \log \left( \frac{p(\mathbf{x})}{p(\mathbf{x}) + q_{\theta}(\mathbf{x})} \right) + q_{\theta}(\mathbf{x}) \log \left( \frac{q_{\theta}(\mathbf{x})}{p(\mathbf{x}) + q_{\theta}(\mathbf{x})} \right) \right], \quad (6)$$

and the total variation distance (TVD),

$$\mathcal{L}^{\text{TVD}}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} |p(\mathbf{x}) - q_{\theta}(\mathbf{x})|. \quad (7)$$

Another example of loss function that can be written in this form is the classical fidelity,

$$\mathcal{L}^{\text{CF}}(\boldsymbol{\theta}) = 1 - \sum_{\mathbf{x} \in \mathcal{X}} \sqrt{p(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{x})}. \quad (8)$$

Notably, any non-data dependent post-processing of an explicit loss retains its explicit character. Thus, any non-data dependent function of an explicit loss (Eq. (2)) may also be considered an explicit loss. For example, the Rényi divergence [69]

$$\mathcal{L}_{R,\alpha}(\boldsymbol{\theta}) = \frac{1}{\alpha - 1} \log \left( \sum_{\mathbf{x}} \frac{p^{\alpha}(\mathbf{x})}{q_{\boldsymbol{\theta}}^{\alpha-1}(\mathbf{x})} \right), \quad (9)$$

with  $0 < \alpha < \infty$  and  $\alpha \neq 1$  can be classified as an explicit loss function.

On the other hand, we define an *implicit loss* as one that can be written as an average over samples drawn from the target and model distributions. That is, an implicit loss function can be expressed as

$$\mathcal{L}_{\text{impl}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_r \sim \{p, q_{\boldsymbol{\theta}}\}} g(\mathbf{x}_1, \dots, \mathbf{x}_r), \quad (10)$$

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

As a key example of an implicit loss, we focus on the commonly used *Maximum Mean Discrepancy* (MMD) [48] loss. The MMD takes the form

$$\mathcal{L}_{\text{MMD}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim q_{\boldsymbol{\theta}}} [K(\mathbf{x}, \mathbf{y})] - 2\mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\theta}}, \mathbf{y} \sim p} [K(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p} [K(\mathbf{x}, \mathbf{y})], \quad (11)$$

where  $K(\mathbf{x}, \mathbf{y})$  is a freely chosen kernel function. We consider the popular choice of a classical *Gaussian kernel*, which is defined as

$$K_{\sigma}(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma}} = \prod_{i=1}^n e^{-\frac{(x_i - y_i)^2}{2\sigma}}. \quad (12)$$

Here,  $\|\cdot\|_2$  is the 2-norm,  $\sigma > 0$  is the so-called *bandwidth* parameter, and  $x_i, y_i$  are the values of bit  $i$  in bitstring  $\mathbf{x}, \mathbf{y}$ , respectively. This kernel in effect provides a continuous measure of the distance between target and model bitstrings.

Interestingly, an implicit loss can always additionally be expressed in a form where it contains the target and model

probabilities. Taking the MMD loss in Eq. (11) as a concrete example, the loss can be re-written as

$$\begin{aligned} \mathcal{L}_{\text{MMD}}(\boldsymbol{\theta}) &= \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})K(\mathbf{x}, \mathbf{y}) \\ &\quad - 2 \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}) \\ &\quad + \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} p(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}). \end{aligned} \quad (13)$$

However, we stress that due to the data-dependence in the kernel  $K(\mathbf{x}, \mathbf{y})$ , the MMD loss function can in general not be classified as an explicit loss.

Nonetheless, this brings us to the subtle point that explicitness and implicitness are in fact not strictly mutually exclusive, i.e., one may be able to find a loss function that satisfies both Eq. (2) and Eq. (10) in specific cases. For example, for the MMD this occurs if the kernel is chosen to be a Kronecker delta function,  $K(\mathbf{x}, \mathbf{y}) = \delta_{\mathbf{x}\mathbf{y}}$ . However, such hybrid losses are very much rare edge cases, and the overwhelming majority of losses are either explicit or implicit. A more detailed discussion of the technical nuances of the explicit and implicit loss distinction is provided in Appendix A.

### C. Loss measurement strategies

Central to the trainability of quantum generative models is the measurement strategy used to estimate the loss. Here we draw a distinction between *conventional and quantum measurement strategies*. For simplicity we now restrict our discussion to implicit quantum generative models such as the QCBM.

The *conventional* measurement strategy, which can be employed by both classical and quantum implicit models, starts by collecting sample data from the target and model distributions in the bases in which the data distribution is modelled, e.g., the computational basis for the case of classical data. For an implicit loss these samples can then be directly used to evaluate the loss function in Eq. (10). For an explicit loss, this is not possible, and instead one needs to use the collected samples to recreate an empirical estimate  $\tilde{q}_{\boldsymbol{\theta}}$  of the true model distributions  $q_{\boldsymbol{\theta}}$ .

More formally, as sketched in Fig. 1, consider the set of bitstrings  $\tilde{Q}_{\boldsymbol{\theta}}$  obtained after collecting  $N$  samples from the model and the empirical model distribution  $\tilde{q}_{\boldsymbol{\theta}}(\mathbf{x})$  constructed from these samples. Then, the statistical estimate of the pairwise explicit loss function  $\tilde{\mathcal{L}}(\boldsymbol{\theta})$  in Eq. (3) can be

expressed as

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in \mathcal{X}} f(\tilde{p}(\mathbf{x}), \tilde{q}_{\boldsymbol{\theta}}(\mathbf{x})). \quad (14)$$

Crucially, since this proxy is all we have access to, the properties of this statistical estimate are what determine the trainability of an explicit loss function when evaluated via the conventional strategy. We note that zero-estimates of the model probabilities with  $\tilde{q}_{\boldsymbol{\theta}}(\mathbf{x}) = 0$  are often ‘clipped’ with a small regularization parameter  $\epsilon \ll 1$  in order to avoid numerical instabilities in the loss computation.

This conventional strategy is somewhat classical in the sense that after sampling is performed on the quantum model, the post processing required to compute the cost is entirely classical. However, ‘more quantum’ measurement strategies are also possible. In this case, a quantum circuit is used to compute functions of the probabilities, potentially more directly and/or collectively.

For example, rather than computing the classical fidelity in Eq. (8) by explicitly computing the probabilities  $q_{\boldsymbol{\theta}}(\mathbf{x})$ , one could encode the target distribution in a quantum state  $|\phi\rangle = \sum_{\mathbf{x}} \sqrt{\tilde{p}(\mathbf{x})} |\mathbf{x}\rangle$  and compute the quantum fidelity

$$\mathcal{L}_{QF}(\boldsymbol{\theta}) := 1 - |\langle \phi | \psi(\boldsymbol{\theta}) \rangle|^2 \quad (15)$$

$$\sim 1 - \left| \sum_{\mathbf{x}} \sqrt{\tilde{p}(\mathbf{x})} q_{\boldsymbol{\theta}}(\mathbf{x}) \right|^2. \quad (16)$$

Up to arbitrary global phase factors (and a mod-square) this is equivalent to the classical fidelity. However, it can be computed via coherent strategies - namely a Loschmidt echo circuit [70–73] or a SWAP test [74, 75]. We note that in this case quantum generative modelling is equivalent to a state learning problem. While this expression seemingly requires the entire training dataset to be loaded into a wavefunction, we present an approach in Sec. III C to estimate this cost using pairwise Hadamard tests.

More generally, it remains an open question if/when commonly encountered losses for generative modelling can be computed using quantum strategies and whether or not this brings any advantages<sup>2</sup>. Nonetheless, we suggest that this is an interesting avenue for future research.

<sup>2</sup> Beyond QCBMs, *Quantum Generative Adversarial Networks* (QGANs) [76] trained with classical discriminators [77–79] in effect use a conventional measurement strategy, whereas their variant with quantum discriminators [80] use a quantum strategy.

## D. Exponential concentration and barren plateaus

For a quantum generative model to be trained successfully, the loss landscape must be sufficiently featured to enable a solution to be found. There is a growing awareness of the importance of barren plateaus, and its sister phenomenon *exponential concentration*, for quantum machine learning [22–29]. A barren plateau (BP) is a loss landscape where the magnitudes of gradients vanish exponentially with growing problem size [22, 24–28, 30–36]. Closely related and equally problematic is exponential concentration where the loss is shown to concentrate with high probability to a single fixed value [23]. This, with high probability, results in poorly trained models using a polynomial number of measurement shots (regardless of the optimization method employed) [26]. More precisely, exponential concentration can be formally defined as follows.

**Definition 1** (Exponential concentration). *Consider a quantity  $X(\boldsymbol{\alpha})$  that depends on a set of variables  $\boldsymbol{\alpha}$  and can be measured from a quantum computer as the expectation of some observable.  $X(\boldsymbol{\alpha})$  is said to be deterministically exponentially concentrated in the number of qubits  $n$  towards a certain fixed value  $\mu$  if*

$$|X(\boldsymbol{\alpha}) - \mu| \leq \beta \in O(1/b^n), \quad (17)$$

for some  $b > 1$  and all  $\boldsymbol{\alpha}$ . Analogously,  $X(\boldsymbol{\alpha})$  is probabilistically exponentially concentrated if

$$\Pr_{\boldsymbol{\alpha}}[|X(\boldsymbol{\alpha}) - \mu| \geq \delta] \leq \frac{\beta}{\delta^2}, \quad \beta \in O(1/b^n), \quad (18)$$

for  $b > 1$ . That is, the probability that  $X(\boldsymbol{\alpha})$  deviates from  $\mu$  by a small amount  $\delta$  is exponentially small for all  $\boldsymbol{\alpha}$ .

A number of causes of exponential concentration and barren plateaus have been identified including using parameterized circuits that are too expressive [22, 24, 30, 42] or too entangling [32, 33, 43]. Hardware noise [34, 35, 81] has also been shown to exponentially flatten the loss landscapes, which strongly hinders the potential of current noisy quantum devices. The exponential concentration can also happen due to randomness in the training dataset [36–38]. In addition, there are studies on the exponential concentration in different QML models including dissipative parametrized quantum circuits [43] as well as quantum kernel-based models [29].

Finally, the choice of loss function can also induce these phenomena. Thus far, loss concentration has predominantly been studied in the context of losses of the form

$$C(\boldsymbol{\theta}) = \text{Tr}[OU(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger], \quad (19)$$

where  $\rho$  is an  $n$ -qubit input state and  $O$  is a Hermitian operator. In particular, it has been shown that ‘global’ [31] losses, i.e., those where  $O$  acts non-trivially on  $\mathcal{O}(n)$  qubits, induce loss concentration even for very shallow random circuits. Conversely, local losses where  $O$  acts non-trivially on at most  $\log(n)$  adjacent qubits (and more generally low-body losses where the adjacency constraint is lifted - see panel a) of Fig. 4) have been shown to enjoy trainability guarantees [31, 39] with shallow unstructured circuits. Furthermore, we note that how barren plateaus affect parametrized quantum circuits with a non-linear loss in the discriminative QML setting has been studied in Ref. [36].

Here we study exponential concentration for generative modelling tasks on classical discrete data using implicit quantum generative models, and use our insights to establish guidelines of how best to train such models. Crucially, in this generative modeling context, the fixed points of the model probabilities tend to be exponentially small and the loss function contains the sum over exponentially many terms. These two together render previously used tools not directly applicable for studying the trainability of quantum generative models.

### III. TRAINABILITY ANALYSIS ON LOSS FUNCTIONS

In this section, we analyse the trainability of different loss functions used in quantum generative modelling.

#### A. Pairwise explicit losses

Part of the power of *quantum* generative models is that they can be used to continuously parameterise and express distributions over discrete data with exponential support. That is, an  $n$ -qubit model can be used to model distributions over  $2^n$  different  $n$ -bitstrings. However, while the true target distribution may have exponential support, the amount of training data  $\tilde{P}$  is in practise restricted. More precisely, for large  $n$  (e.g.,  $n > 50$ ), it is reasonable to assume that the number of bitstrings in the training dataset scales at most polynomially in  $n$ . Similarly, the number of bitstrings samples obtained from the model must also scale at most polynomially in  $n$ . That is,  $|\tilde{P}|, |\tilde{Q}_\theta| \in \mathcal{O}(\text{poly}(n))$ .

This discrepancy between the polynomial support of the training data and the exponential support of the model, can make it highly challenging to train implicit models using pairwise explicit loss functions. In loose terms, the problem is that the only bitstrings that contribute to the evaluation of a statistical estimate of an explicit cost are

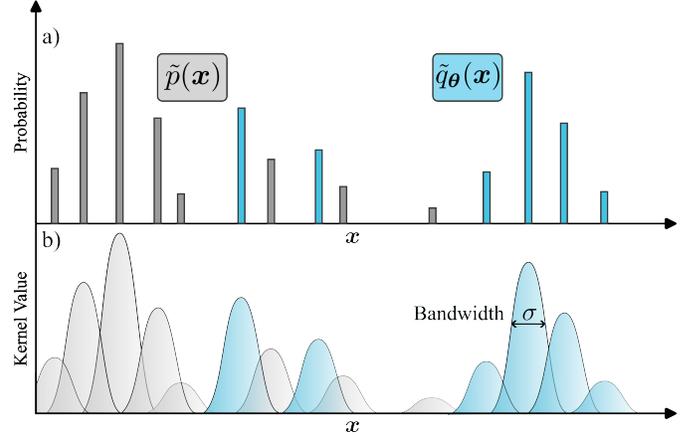


Figure 2. **The problem with pairwise explicit losses.** In the space with  $2^n$  unique  $n$ -bit bitstrings, samples  $\mathbf{x}$  generated from an uninformed model with high probability do not coincide with any of the training bitstrings. In other words, the empirical model distribution  $\tilde{q}_\theta(\mathbf{x})$  and the training distribution  $\tilde{p}(\mathbf{x})$  do not both have non-zero probabilities for any bitstring  $\mathbf{x}$ . On the other hand, an implicit loss function such as the MMD provides a continuous measure of distance between the distributions by use of a Gaussian kernel with bandwidth  $\sigma$ .

those corresponding to bitstrings  $\tilde{P}$  in the training data. To estimate the loss one thus needs good estimates of the model distributions over the support of  $\tilde{P}$ . However, for an implicit model these estimates are obtained via sampling and the set  $\tilde{P}$  contains an exponentially small proportion of the total number of bitstrings. As such, for *generic models* (we will explain what we mean by generic below), the probability of measuring any bitstring in the training set will also be exponentially small (as sketched in Fig. 2), leading to a poor statistical estimate of the loss.

#### 1. Concentration of Pairwise Explicit Losses

To make this line of argument more concrete, the first family of models we will consider are those where the individual model probabilities  $q_\theta(\mathbf{x})$  are exponentially concentrated over different values of  $\theta$ . This is the case for a large family of unstructured parameterised quantum circuits. Since estimating  $q_\theta(\mathbf{x})$  is equivalent to computing the expectation value of the global projector  $|\mathbf{x}\rangle\langle\mathbf{x}|$ , the concentration of  $q_\theta(\mathbf{x})$  can be viewed as resulting from the global-measurement induced barren plateau phenomenon [31]. In this case, concentration is observed even for an ansatz that is comprised of only a single layer of single-qubit rotations. However, alternative phenomena (e.g. noise [34] or expressibility [30]) can also lead to the exponential concentration

of  $q_{\theta}(\mathbf{x})$ . More formally, the following proposition holds.

**Proposition 1** (Concentration of model). *For all possible bitstrings  $\mathbf{x} \in \mathcal{X}$ , the underlying probability  $q_{\theta}(\mathbf{x})$  of the quantum model exponentially concentrates towards some exponentially small fixed point  $\mu \in O(1/b^n)$  for  $b > 1$  if the quantum generative model is constructed with:*

- A single layer of random single qubit gates  $U(\theta) = \bigotimes_{i=1}^n U_i(\theta_i)$ . Or, more precisely, if  $\{U_i(\theta_i)\}_{\theta_i}$  forms a local 2-design on qubit  $i$  [31].
- $L$  layers of random  $k$ -local 2-designs, i.e.,  $U(\theta) = \prod_{l=1}^L \bigotimes_{j=1}^{n/k} U_{l,j}(\theta_{l,j})$  with each  $U_{l,j}(\theta_{l,j})$  acting on  $k$  qubits and  $\{U_{l,j}(\theta_{l,j})\}_{\theta_{l,j}}$  forming a  $k$ -local 2-design over  $\theta_{l,j}$  [31].
- A parameterised quantum circuit  $U(\theta)$  such that its ensemble over  $\theta$  i.e.,  $\{U(\theta)\}_{\theta}$  forms an approximate 2-design on  $n$  qubits [22, 30]. This holds even for the problem-inspired circuits [24].
- A linear-depth quantum circuit subject to local Pauli noise between each layer [34].

Proposition 1 provides examples of cases where the model probabilities exponentially concentrate over *all* bitstrings in  $\mathcal{X}$ . However, we find that in fact trainability difficulties arise even if model probabilities are only exponentially concentrated over the training dataset (but perhaps not on points outside the dataset). That is, all that is required for untrainability is that the probability of measuring a sample that is also in the dataset is practically zero. This is likely to be the case even for highly structured quantum circuits if the generative model is built without a strong inductive bias. We formalise this intuition in Appendix B 2.

We now argue that the exponential concentration of probabilities  $q_{\theta}(\mathbf{x})$  over the dataset causes  $\tilde{\mathcal{L}}(\theta)$  to also exponentially concentrate. To understand why, let us look at the probability of measuring one specific bitstring (e.g.,  $\mathbf{x}_0$  - the all-zero bitstring) and assume that  $q_{\theta}(\mathbf{x}_0)$  is exponentially concentrated towards some exponentially small value  $\mu$ . Then, for any given parameter constellation, it is highly likely that  $q_{\theta}(\mathbf{x}_0)$  is exponentially close to  $\mu$ . To estimate  $q_{\theta}(\mathbf{x}_0)$  on a quantum computer we sample  $N$  bitstrings from the quantum model and record the observations. The chance that none of the sampled bitstrings are the specific bitstring that we are interested in is  $(1 - q_{\theta}(\mathbf{x}_0))^N \approx 1 - N\mu$ . However, the number of circuits  $N$  that can be efficiently run is necessarily limited - here we will assume  $N \in \text{poly}(n)$ . Thus we have that the probability of not measuring the bitstring we are interested in is exponentially close to 1. That is, the statistical estimate of

$\tilde{q}_{\theta}(\mathbf{x}_0)$  is almost always zero. We can then generalize this intuition for a single bitstring to the estimation of each of the (polynomially many) target bitstrings and therefore the whole loss function. The following theorem formalizes this argument.

**Theorem 1** (Concentration of pairwise explicit loss for concentrated models). *Consider the loss function of the form in Eq. (3). 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*

$$\mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta}) = \sum_{\mathbf{x} \in \tilde{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \tilde{Q}_{\theta}} f(0, \tilde{q}_{\theta}(\mathbf{x})), \quad (20)$$

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.,

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta})] \geq 1 - \delta, \quad (21)$$

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

As a direct consequence of Theorem 1, the following corollary gives the concentration points of some specific explicit loss functions mentioned in this work.

**Corollary 1** (Concentration points of common explicit loss functions). *Under the same conditions as in Theorem 1, the following loss functions concentrate at*

- *KL-divergence:*

$$\mathcal{L}_0^{\text{KLD}}(\tilde{P}, \tilde{Q}_{\theta}) = \sum_{\mathbf{x} \in \tilde{P}} \tilde{p}(\mathbf{x}) \log\left(\frac{\tilde{p}(\mathbf{x})}{\epsilon}\right). \quad (22)$$

Here  $\epsilon \ll 1$  is a clipping value, which is common practice to avoid the singularity of the logarithm at  $q_{\theta}(\mathbf{x}) = 0$ .

- *Classical fidelity:*

$$\mathcal{L}_0^{\text{CF}}(\tilde{P}, \tilde{Q}_{\theta}) = 1. \quad (23)$$

- *Reverse KL-divergence:*

$$\mathcal{L}_0^{\text{rev-KLD}}(\tilde{P}, \tilde{Q}_{\theta}) = \sum_{\mathbf{x} \in \tilde{Q}_{\theta}} \tilde{q}_{\theta}(\mathbf{x}) \log\left(\frac{\tilde{q}_{\theta}(\mathbf{x})}{\epsilon}\right). \quad (24)$$

- *Total variation distance:*

$$\mathcal{L}_0^{\text{TVD}}(\tilde{P}, \tilde{Q}_\theta) = 2. \quad (25)$$

Looking at the expressions for the fixed points given above, in the case of the KL divergence, classical fidelity and total variational distance, the fixed point is independent of  $\theta$ . Thus it is clear that the costs cannot be used to train the quantum circuit model. In the case of the reverse KL divergence, the fixed point depends on  $\theta$  but is independent of the training data and thus the reverse KL also cannot be used to train the model to learn the target distribution.

More generally, for all explicit losses of the form Eq. (3), the concentration point  $\mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)$ , Eq. (20), can be separated into two terms: (i) the term that involves only  $\tilde{P}$  and (ii) the other that involves only  $\tilde{Q}_\theta$ . In other words, the  $\theta$  dependence of the estimator of the loss is independent of the target distribution and thus the estimate of the loss is worthless for training the generative model. This no-go result is rigorously established in Corollary 2. Our approach is to show that the loss function at two arbitrary parameter values  $\theta_1$  and  $\theta_2$ , contains no information about the training distribution.

**Corollary 2** (Untrainability of pairwise explicit loss functions). *Under the same conditions as in Theorem 1, the probability that the difference between the two statistical estimates of the loss function at  $\theta_1$  and  $\theta_2$  does not contain any information about the training distribution is exponentially close to 1. Particularly, we have*

$$\Pr_{\tilde{Q}_\theta, \theta}[\tilde{\mathcal{L}}(\theta_1) - \tilde{\mathcal{L}}(\theta_2) = \Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2})] \geq 1 - 2\delta, \quad (26)$$

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  $\theta_1$  (and  $\theta_2$ ), as well as

$$\Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2}) = \sum_{\mathbf{x} \in \mathcal{Q}_{\theta_1}} f(0, \tilde{q}_{\theta_1}(\mathbf{x})) - \sum_{\mathbf{x} \in \mathcal{Q}_{\theta_2}} f(0, \tilde{q}_{\theta_2}(\mathbf{x})), \quad (27)$$

with  $\mathcal{Q}_{\theta_1}$  (and  $\mathcal{Q}_{\theta_2}$ ) being a set of unique bit-strings 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  $\tilde{p}(\mathbf{x}) \in \tilde{P}$ .

To support our analytic claims we further conducted a numerical study of the exponential concentration of pairwise explicit costs. For concreteness, we here decided to focus on the KL divergence. In Fig. 3, we plot the mean and variance (over  $\theta$ ) of the KL divergence for the target distribution  $\tilde{p}(\mathbf{0}) = 1$  as a function of the number of

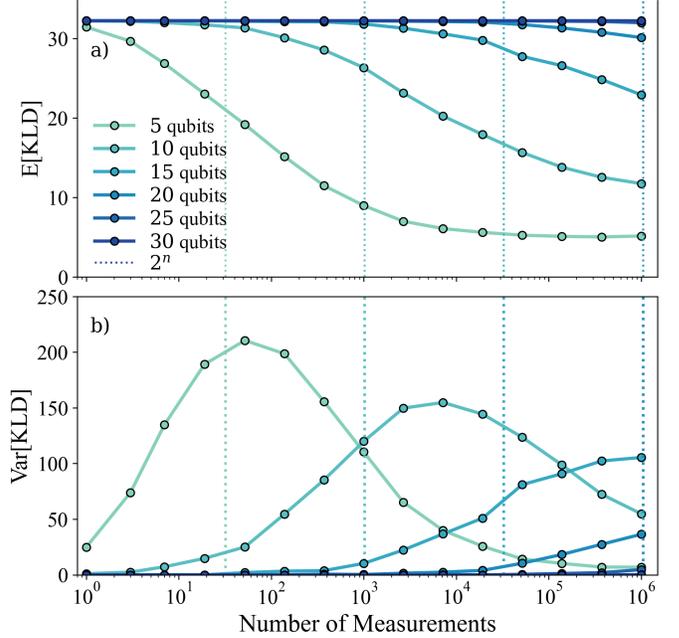


Figure 3. **Variance of the KL divergence with finite shots.** Concentration of the KL divergence loss as a function of the number of measurements and qubits for random product state circuits. Here we take the target distribution to be  $\tilde{p}(\mathbf{0}) = 1$  and take the cutoff of the KLD to be  $\epsilon = 10^{-14}$ . Vertical lines indicate where the number of measurements equal  $2^n$ . Thus, we see that the the KLD estimate is biased upwards with any finite number of measurements, and the number of measurements required to achieve a reasonable level of uncertainty increases exponentially with the number of qubits  $n$ .

measurement shots and qubits. For simplicity we take our model to be a (Haar) random product state.

We see in Fig. 3a) that with a polynomial number of measurements, as per Eq. (22), the empirical estimate of the loss concentrates at  $\log(1/\epsilon) \approx 32.2$  for  $\epsilon = 10^{-14}$ . Correspondingly, with a polynomial number of measurements the variance in Fig. 3b) is exponentially close to zero. Using an exponential number of measurements, the estimate of the KL tends towards its true value and the variance is again small. The transition between these two regimes is marked by a very high variance corresponding to the case where the measurement count is high enough for there to be some overlap between the sampled bits strings and the  $\mathbf{0}$  bitstring, but not enough overlap to obtain a reliable estimate of  $q_\theta(\mathbf{0})$ . This results in the loss estimate to sporadically fluctuate between  $\log(1/\epsilon)$  and  $\log(1/q_\theta(\mathbf{x}))$  with  $q_\theta(\mathbf{x}) > 0$ . While in Fig. 3 the target dataset consists of a single bitstring, larger datasets only shift the curves to the

left by a polynomial amount.

*Broader Implications.* While our results above are formulated for training QCBMs with pairwise explicit costs, we argue that the underlying problem is more general and immune to simple solutions. One approach, for example, might be to take non-data-dependent functions of pairwise explicit losses, as in the case of the Rényi-divergence in Eq. (9). However, such loss functions exponentially concentrate in the same manner as the explicit losses themselves when employing the conventional measurement strategy. A more promising but challenging approach would be to attempt to measure such losses via quantum strategies. We discuss this further in Section III C.

More generally, while we provide strict no-go results only for pairwise explicit losses, we believe that any explicit losses in the general form of Eq. (2) will suffer from concentration or exponential imprecision due to the inherent inability of implicit models to accurately estimate the model probabilities in polynomial time<sup>3</sup>. We are however not aware of any practical explicit loss function that cannot be brought into the pairwise explicit form.

We further stress that our results hold for unstructured ansätze or ansätze that lack an appropriate inductive initial bias. Thus, while explicit losses such as the KLD will not work at scale with implicit models straight out-of-the-box, our no-go theorems could be side-stepped using clever initialization strategies in conjunction with specialized ansätze. For example, while we argue in Appendix B 2 that initializing the quantum circuit model on a subset of training states will not alleviate the fundamental issue when using a generic ansatz, this may work if one leverages a quantum circuit that constrains the model to the symmetry sector of the data. Among other hard constraints, this is conceivable if the data consists only of samples with a certain hamming weight or cardinality, as it can be the case in certain financial applications [82, 83]. However, many real world datasets may not contain strong symmetries that one can leverage so straight-forwardly. It is therefore critically important to study the effect of strong parameter initializations and inductive biases using explicit losses— both theoretically and experimentally.

## B. Implicit losses: Maximum Mean Discrepancy

In the previous section we saw that an explicit loss function, used in conjunction with an implicit generative model

and the conventional sampling strategy, exhibits exponential concentration and hence is untrainable. The root cause was, at least in part, a miss-match between using an explicit loss function with an implicit model. Thus it is natural to ask whether an implicit quantum loss would fare better.

Here we focus on analysing the MMD loss function (see Eqs. (11) and (13)), which is a commonly-used implicit loss. In contrast to the pairwise explicit losses discussed previously, each bitstring drawn from the model is generally compared with all training bitstrings, with the kernel function  $K(\mathbf{x}, \mathbf{y})$  controlling the contribution of each comparison. With a poor choice in kernel it is clear that the MMD will be susceptible to exponential concentration. For example, the Gaussian kernel with the bandwidth  $\sigma \rightarrow 0$  is equivalent to a delta function kernel,  $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle = \delta_{\mathbf{x}\mathbf{y}}$ . In this case the MMD reduces to the pairwise explicit loss  $\sum_{\mathbf{x} \in \mathcal{X}} (p(\mathbf{x}) - q_{\theta}(\mathbf{x}))^2$  (see Appendix A for details), and consequently is subject to our no-go result in Theorem 1. This thus prompts the question of how exactly  $\sigma$  affects trainability.

*Properties of the MMD loss.* To study the properties of the MMD loss, it is helpful to note that each term in the MMD can be viewed as the expectation value of an observable whose properties depend on the choice of  $\sigma$ . This change in perspective allows us to leverage existing knowledge from the VQA trainability literature. In particular, prior no-go results on VQAs with observable-type loss functions are now directly applicable here, including those on cost function induced [31], expressibility-induced [22, 30], and noise-induced [34] barren plateaus.

Specifically, each term in the MMD can be written as

$$\mathcal{M}(\rho, \rho') = \text{Tr} \left[ O_{\text{MMD}}^{(\sigma)} (\rho \otimes \rho') \right], \quad (28)$$

where we have defined the MMD observable

$$O_{\text{MMD}}^{(\sigma)} := \sum_{\mathbf{x}, \mathbf{y}} K_{\sigma}(\mathbf{x}, \mathbf{y}) |\mathbf{x}\rangle \langle \mathbf{x}| \otimes |\mathbf{y}\rangle \langle \mathbf{y}|. \quad (29)$$

This observable acts on  $2n$  qubits, namely  $n$  qubits corresponding to the QCBM,  $\rho_{\theta} = |\psi(\theta)\rangle \langle \psi(\theta)|$ , and  $n$  qubits corresponding to the dataset,  $\rho_{\bar{p}} = \sum_{\mathbf{y}} \tilde{p}(\mathbf{y}) |\mathbf{y}\rangle \langle \mathbf{y}|$ . For the first term in the MMD, both  $\mathbf{x}$  and  $\mathbf{y}$  are sampled from the QCBM and we have  $\rho = \rho' = \rho_{\theta}$ . The cross-term instead has  $\rho = \rho_{\theta}$  and  $\rho' = \rho_{\bar{p}}$ , and the final term has  $\rho = \rho' = \rho_{\bar{p}}$ .

In the Pauli basis, the MMD observable  $O_{\text{MMD}}^{(\sigma)}$  takes the elegant form

$$O_{\text{MMD}}^{(\sigma)} = \sum_{l=0}^n w_{\sigma}(l) D_{2l}, \quad (30)$$

<sup>3</sup> A possible exception is if a particular implicit model instead allows for efficient estimation of gradients of an explicit loss function, as it is the case for RBMs training on the KL divergence loss function

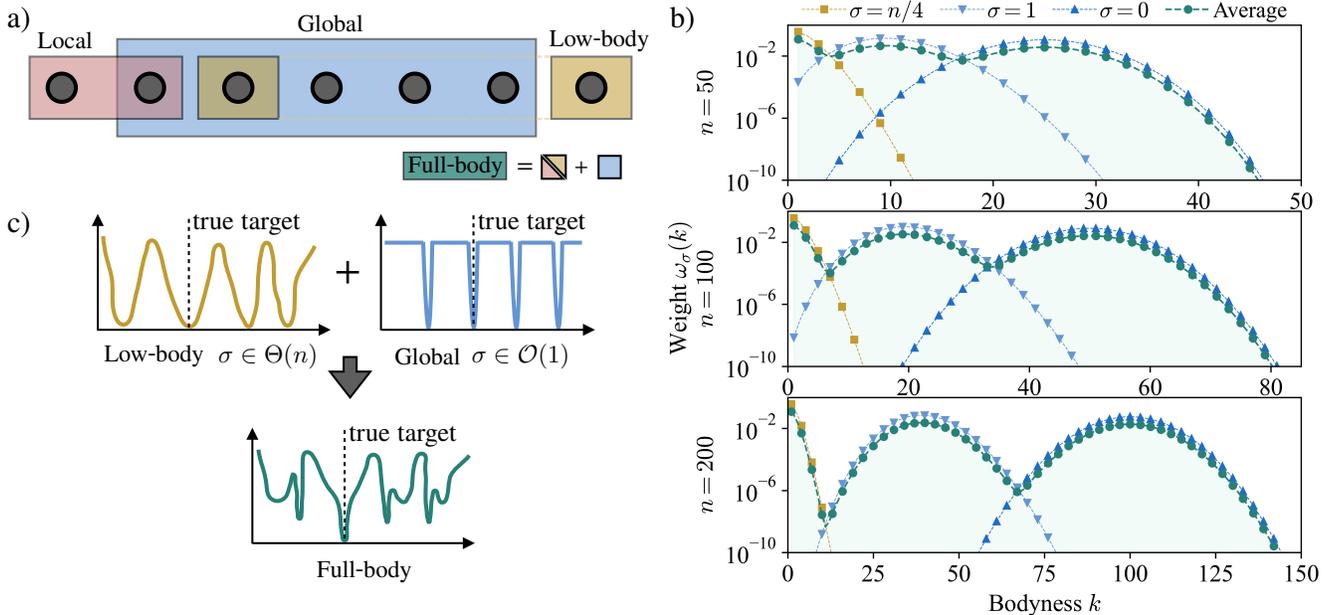


Figure 4. **Bodyness of the MMD loss.** a) We illustrate the difference between ‘low-body’, ‘local’, ‘global’ and ‘full-body’ operators. An operator  $O$  is low-bodied if it acts non-trivially on at most  $\mathcal{O}(\log(n))$  qubits. If a low-bodied operator acts on qubits that are adjacent to each other, then  $O$  is said to be local. On the other hand, if it acts non-trivially on  $\Theta(n)$  qubits. Lastly, a full-body operator consists of the sum of several operators that are low-body/local and global. b) We depict the weight  $w_\sigma(l)$  for the terms in the MMD operator as a function of their bodyness  $k$  for  $n = 50, 100$  and  $200$  qubits and the bandwidths  $\sigma = 0, 1$  and  $n/4$ . The average weight over these three  $\sigma$  values is also shown. For small  $\sigma$ , the MMD operator is a sum of predominantly global operators, i.e., with  $\sigma \in \mathcal{O}(1)$  the mean bodyness is  $\Theta(n)$ . In contrast,  $\sigma \in \Theta(n)$  results in predominantly low-bodied operators. c) Sketch of the expected landscapes for low-body, global and full-body losses respectively. Because low-body and global operators are exclusively sensitive to low-body and global features, respectively, their loss landscapes exhibit spurious minima, which don’t coincide with the minimum of the true target distribution. A full-body loss on the other hand should have a single optimal solution where all its constituent operator’s minima align.

where  $D_{2l}$  are normalized  $2l$ -body diagonal operators (defined explicitly in Appendix C 1), and

$$w_\sigma(l) = \binom{n}{l} (1 - p_\sigma)^{n-l} p_\sigma^l \quad (31)$$

are Bernoulli-distributed weights with effective probability

$$p_\sigma = (1 - e^{-1/2\sigma})/2. \quad (32)$$

Thus estimating the MMD loss function in Eq. (11) using a batch of measurements  $\tilde{Q}$  is equivalent to using the same measurements to estimate a weighted expectation of the observables  $D_{2l}$ .

The properties of the MMD observable clearly depend on the distribution of the terms of different bodyness through the  $w_\sigma(l)$  factor. Fig. 4 shows how  $w_\sigma(l)$  are distributed for different  $\sigma$ . Owing to the Bernoulli-distributed weights, we can straight-forwardly provide the average bodyness of

$O_{\text{MMD}}^{(\sigma)}$ , which is given by

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] = 2np_\sigma, \quad (33)$$

and the variance in the bodyness, which is

$$\text{Var}_{l \sim w_\sigma(l)}[2l] = 4np_\sigma(1 - p_\sigma). \quad (34)$$

From these expressions it follows that the MMD loss is predominantly composed of global operators when  $\sigma \in \mathcal{O}(1)$ . More concretely the following proposition holds.

**Proposition 2** (MMD consists largely of global terms for  $\sigma \in \mathcal{O}(1)$ ). *For  $\sigma \in \mathcal{O}(1)$ , the average bodyness of the MMD operator containing Pauli terms with weight  $w_\sigma(l)$  is*

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] \in \Theta(n). \quad (35)$$

Similarly, the variance in the bodyness is given by

$$\text{Var}_{l \sim w_\sigma(l)}[2l] \in \Theta(n). \quad (36)$$

This shows that with fixed-size bandwidths  $\sigma$ , as is commonly done (e.g., Ref. [51]), the MMD suffers from global loss function-induced barren plateaus [31] and hence is untrainable. This practice of using constant bandwidths is carried over from classical ML literature [84–86], but Proposition 2 shows that this is fundamentally incompatible with quantum generative models using unstructured circuits.

In contrast, we show that if the bandwidth scales linearly in the number of qubits,  $\sigma \in \Theta(n)$ , the MMD loss function is approximately low-bodied. We recall that being low-bodied is more general than being *local*, the latter corresponding to the case where an operator is low-bodied and each term only acts non-trivially on adjacent qubits. The following proposition formalizes this relation by quantifying the error made when truncating the MMD observable after a certain bodyness.

**Proposition 3** (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  $O_{\text{MMD}}^{(\sigma)}$ ,*

$$\tilde{O}_{\text{MMD}}^{(\sigma,k)} := \sum_{l=0}^k w_{\sigma}(l) D_{2l}, \quad (37)$$

where  $w_{\sigma}(l)$  are Bernoulli-distributed weights defined in Eq. (31). For  $\sigma \in \Theta(n)$ , the difference between the exact and local approximation of the loss is bounded as

$$|\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\boldsymbol{\theta})| \leq \epsilon(k), \quad (38)$$

with

$$\epsilon(k) \in \mathcal{O}(n(c/k)^k), \quad (39)$$

for some positive constant  $c$ .

This implies that one can view the MMD loss with a bandwidth  $\sigma \in \Theta(n)$  as composed almost exclusively of low-body contributions. We therefore expect, given the results of Refs. [31, 39], that the MMD is trainable for  $\sigma \in \Theta(n)$  for quantum generative models which employ shallow quantum circuits. We note that there appears to be no merit in increasing  $\sigma$  beyond  $\Theta(n)$ , as that simply increases the relative weight of the constant  $l = 0$  term in Eq. (30). That is, the MMD operator tends towards the trivial identity measurement for  $\sigma \rightarrow \infty$ .

To probe this further, and get a better understanding of the effect of  $\sigma$  on the trainability of the MMD loss, we start by considering the case of QCBM with a product ansatz. This allows us to find a closed-form expression of

the MMD variance as a function of the circuit parameters (Supplemental Proposition 2) from which we can study the concentration of the MMD for different  $\sigma$  values. Our findings are summarized by the following Theorem (proven in Appendix C 2c).

**Theorem 2** (Product ansatz trainability of MMD, informal). *Consider the MMD loss function  $\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})$  as defined in Eq. (11), which uses the classical Gaussian kernel as defined in Eq. (12) 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  $\sigma$ .*

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

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] \in \mathcal{O}(1/b^n), \quad (40)$$

with some  $b > 1$ .

On the other hand, for  $\sigma \in \Theta(n)$ , we have

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] \in \Omega(1/n). \quad (41)$$

We numerically verify Theorem 2 in Fig. 5. In panel a) we show that the analytical predictions for different bandwidths coincide perfectly with the numerical estimates. The exponentially vanishing loss variances observed for  $\sigma \in \mathcal{O}(1)$  are expected to render the loss untrainable. This is demonstrated in panel b), where we further train a QCBM with  $\sigma = n/4$  (which approximately maximizes the variance) and  $\sigma = 1$ . We find that a QCBM with  $\sigma = n/4$  can be successfully trained even for  $n = 1000$  qubits. In contrast, the training starts to fail to learn the  $|\mathbf{0}\rangle$  target state after  $n \approx 50$  and is fully untrainable at  $n = 100$  when  $\sigma = 1$  is used.

It is interesting to note that the approximately optimal bandwidth  $\sigma \sim \frac{n}{4}$  for the product state ansatz coincides with the so-called *median heuristic* [48] from classical ML literature. For random circuits, the median (hamming) distance between bitstrings is in fact  $\frac{n}{2}$ , which we satisfy with the factor of 2 in our kernel convention.

To go towards more practical generative modelling, we recall that Ref. [39] proves that cost functions of the form of Eq. (19) using  $2k$ -body observables with  $k \in \mathcal{O}(\log(n))$  are trainable using 1D-random  $\log(n)$  depth circuits. Since Proposition 3 implies that the MMD is well approximated by a  $\log(n)$ -body cost, it should follow that the MMD is also trainable at  $\log(n)$  depths. There are a few technical caveats associated with constructing a full proof. For example, the first term of the MMD requires working with

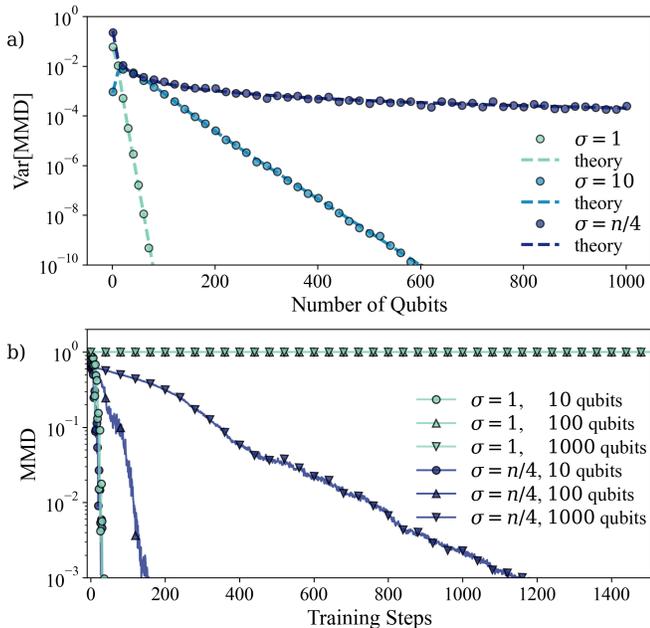


Figure 5.  $\sigma$ -dependence of the MMD loss function. a) Comparison of the MMD variance between the analytical prediction in Eq. (C84) and empirical variance using 100 measurements from a product state ansatz. b) Training a product state ansatz on the  $|0\rangle$  target state for  $\sigma = 1$  and  $\sigma = n/4$  using the CMA-ES [87] optimizer and 512 measurements. In both cases the QCBM ansatz consists of a single layer of Ry rotations on each qubit.

4-designs instead of 2-designs, and the second term depends on the target distribution, leading to additional subtleties. However, there is no strong reason to expect that these technicalities make the MMD untrainable. Hence, we propose the following conjecture.

**Conjecture 1.** *A QCBM composed of a shallow  $\log(n)$  depth unstructured circuit is trainable with the MMD loss function as defined in Eq. (11) using the Gaussian kernel in Eq. (12) as long as the bandwidth  $\sigma \in \Theta(n)$ . That is, we have*

$$\text{Var}_{\theta}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\theta)] \in \Omega(1/\text{poly}(n)). \quad (42)$$

This conjecture is further supported by our numerical evidence for the trainability of the MMD for deeper circuits and more realistic datasets shown in Fig. 6. Here we plot the loss variance as a function of circuit depth  $L$  and the number of qubits  $n$  for  $\sigma = n/4$  on four datasets from four different target distributions. We observe that the polynomial scaling of the loss variance does in fact extend beyond product states to shallow circuits, i.e.,  $L \in \mathcal{O}(\log(n))$ .

However, for sufficiently deep circuits, i.e.,  $L \in \Omega(n)$ , the MMD variance appears to decay exponentially. This aligns with expressibility-induced barren plateaus observed in other VQA applications, which occur even for maximally local loss functions, i.e.,  $k = 1$ .

*Large gradients are not enough.* Our results so far appear to indicate that picking a single bandwidth  $\sigma \in \Theta(n)$  maximizes the trainability of the MMD loss function with a Gaussian kernel. While it is true that this choice maximizes the expected magnitude of initial gradients for a QCBM, non-vanishing gradients are a necessary condition but not sufficient to guarantee reliable training performance. And in fact it turns out that while low-body losses exhibit large gradients they come with other limitations. Particularly, we show that the bodyness of a generative loss function defines the maximal order of marginals of the target distribution that can be distinguished. That is, the model only learns to match the target distribution on subsets of bits, i.e. on its marginals. This introduces a continuous family of minima which are indistinguishable from the true minimum when using a low-bodied loss function, but which are systematically wrong for the purposes of generative modelling. The worry is that the non-vanishing loss gradients in low-bodied losses are predominantly due to the presence of such spurious minima and do not point in the direction of the true global minimum. This is sketched in Fig. 4.

Formally, let  $q_{\theta}(\mathbf{x}_A)$  denote the marginal model distribution on a subset  $A \subseteq \{1, 2, \dots, n\}$  of qubits, and  $\tilde{p}(\mathbf{x}_A)$  the marginal target distribution on that same subset. For more details we refer to Eq. (C139) and Eq. (C141) in Appendix C3. The connection between the bodyness of the loss operator and the marginals of the model and target distributions is then formalized in the following Proposition.

**Proposition 4** (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*

$$q_{\theta}(\mathbf{x}_A) = \tilde{p}(\mathbf{x}_A), \quad (43)$$

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

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma, k)}(\theta) = 0. \quad (44)$$

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

$$q_{\theta}(\mathbf{x}_B) \neq \tilde{p}(\mathbf{x}_B). \quad (45)$$

In other words, if the MMD operator can be approximated well by a truncated operator with at most  $2k$ -body

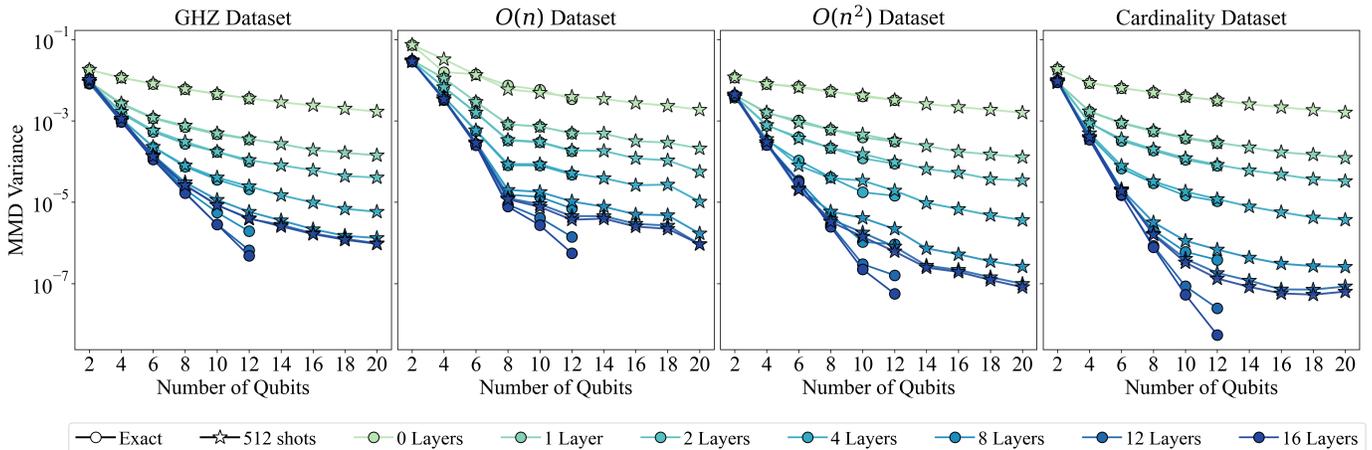


Figure 6. **Study of loss concentration with the MMD loss function.** Numerical evidence that the MMD loss with Gaussian bandwidth parameter  $\sigma = n/4$  does not exhibit global or explicit loss function barren plateaus, but does exhibit loss concentration with deep quantum circuits. We study the loss concentration in randomly initialized line-topology circuits for various datasets, and increasing number of qubits  $n$  and circuit depth. The GHZ dataset consists of the all-0 and all-1 bitstrings ( $\mathcal{O}(1)$  support), the  $\mathcal{O}(n)$  and  $\mathcal{O}(n^2)$  datasets consist of  $n$  and  $n^2$  random bitstrings, respectively, and the cardinality dataset contains all bitstrings with  $\frac{n}{2}$  cardinality ( $\mathcal{O}(2^n)$  support). There does not appear to be a strong data-dependence for the magnitude of the loss variance.

terms, model distributions that match the target distribution exactly up to  $k$ -body marginals or higher cannot be distinguished from ones that match fully. As an example of such distributions, consider the uniform distribution over the bitstrings  $[001, 011, 101, 110]$ , where the third bit is the bit-wise addition of the previous two bits. Using only second-order marginals, it is not possible to distinguish this correlated distribution from the uniform distribution over all eight possible outcomes.

Notably, long-range correlations in the data can still be learned by the low-bodied MMD loss, just not ones that are particularly high-order<sup>4</sup>. Not all distributions will however exhibit such higher-order correlations and thus some distributions will be learnable using losses composed of low-body terms.

Proposition 4 thus establishes that to fulfil the promise of quantum generative models, that is to be able to learn both long-range *and* many-body correlations, one cannot use exclusively low-body losses. However, such a requirement is in immediate tension with the low-bodiedness required for the trainability guarantees (see Theorem 2). In particular, in Proposition 3 we show that for  $\sigma \in \Theta(n)$  the contribution of  $k \in \Theta(n)$  terms are exponentially small in

$n$ . Thus, although the loss is still strictly faithful given an infinite shot budget, with a reasonable shot budget we will not be able to resolve the contribution from the exponentially small high-body terms. Hence, there can be spurious minima that we cannot resolve from the true minimum and therefore for all practical purposes the loss is effectively not truly faithful.

One approach to resolving this tension would be to adapt the initial value of  $\sigma$  from  $\Theta(n)$ , where the loss exhibits large gradients but predominantly learns low-order marginals, towards  $\mathcal{O}(1)$  to also learn high-order correlations as the model improves. This is in line with studies from the classical ML literature showing that bandwidths for optimal MMD performance are oftentimes smaller than the so-called median heuristic [88–90], which coincides with our result of  $\sigma \in \Theta(n)$ . Another approach, which is also already employed in classical ML literature, is to use a kernel that averages the effects of several  $\sigma$  [84–86]. That is, the kernel is taken to be

$$K_{\mathbf{c}}(\mathbf{x}, \mathbf{y}) = \frac{1}{|\mathbf{c}|} \sum_{i \in \mathbf{c}} K_{\sigma_i}(\mathbf{x}, \mathbf{y}) \equiv \sum_{l=1}^k \langle w_{\sigma}(l) \rangle_{\mathbf{c}} D_{2l} \quad (46)$$

for a set of bandwidths  $\mathbf{c} = \{\sigma_1, \sigma_2, \dots\}$ . The resulting weight of each  $2k$ -body term of the new MMD observable is an average of the weightings corresponding to each  $\sigma_i$  in  $\mathbf{c}$  as shown in Fig. 4. Theorem 2 shows that for a QCBM without inductive bias to not fall prey to exponential concentration, at least one of the  $\sigma_i$  needs to be  $\Theta(n)$ . But the

<sup>4</sup> Note that this is in contrast to a loss composed purely of local terms which would be restricted to learning local/short-range correlations.

results of Proposition 4 suggest that for data sets exhibiting high-order correlations a small bandwidth  $\sigma_i \in \mathcal{O}(1)$  is required for correct convergence. It stands to reason that the optimal set  $\mathbf{c}$  contains a spectrum of bandwidths that both enable trainability and faithful convergence to the target distribution (as sketched in Fig. 4c). How successful this strategy is in practice remains to be determined.

*Broader Implications.* Our work highlights that one can treat classical machine learning losses as quantum observables to study their properties. This implies that our results transfer to other types of quantum generative models beyond the QCBM that will also be affected by the fundamental limitations described by Proposition 4. In fact, we show in Appendix C4 that any generative modelling loss function for classical data that can be brought into the form  $\mathcal{L}(\boldsymbol{\theta}) = \text{Tr}[\mathcal{M}\rho_{\boldsymbol{\theta}}]$ , with a diagonal measurement operator  $\mathcal{M}$ , faces the same tension described above. That is, if  $\mathcal{M}$  contains at most  $k$ -body terms in the Pauli basis representation, then the loss cannot distinguish two distributions that agree on all  $k$ -order marginals but disagree on higher-order marginals. Thus losses composed exclusively of local terms (with the conventional measurement strategy) cannot be used in generative modelling to learn complex higher-order correlations.

With a little thought it becomes clear that an exclusively global loss is also undesirable. Not only do such losses exhibit exponential concentration for unstructured circuits, they will also in general possess spurious minima in virtue of only probing global properties of the distribution (i.e. the average global parity), as shown in Fig 4. Instead we advocate using *full*-body losses which contain both low and high-body terms, such as those obtained by averaging in Fig. 4. Even then, global contributions cannot be vanishingly small or else they will not be possible to resolve with a realistic shot budget.

For another example, one may aim to train a quantum generative model using a QGAN framework, where a Discriminator  $D$  provides a score  $D(\mathbf{x})$  to every sample. The corresponding operator can then be written as  $\mathcal{M} = \sum_{\mathbf{x}} D(\mathbf{x})|\mathbf{x}\rangle\langle\mathbf{x}|$ . The Discriminator may have to initially implement an effectively low-bodied operator to facilitate initial gradients, but later in training become higher-bodied to learn global features. That is not to say that the Discriminator should only classify marginals of the bit-string such as in Ref. [91]. Rather, the architecture and initialization should be such that the operator  $\mathcal{M}$  in the Pauli basis initially contains low-body terms but can include high-body terms during convergence. Interestingly, the interpolation from trainable to faithful could be naturally full-filled during training when the Generator and Discriminator are optimized in tandem.

Fine tuning the interplay between the loss function gradients, density of local minima and the faithfulness of a generative loss is beyond the scope of this work, but is an important direction for future research. We especially emphasize the necessity to evaluate the implications of our results on models and datasets of practical relevance. In Section IV we take steps in this direction by investigating training a QCBM to model real data from the HEP domain.

### C. Quantum strategies: quantum fidelity

While the classical fidelity in Eq. (8) is an explicit cost function, the quantum fidelity, defined in Eq. (15), allows for a simple known quantum estimation strategy. Key to the quantum fidelity loss is to interpret the training distribution as a target state  $|\phi\rangle = \sum_{\mathbf{x}} \sqrt{\tilde{p}(\mathbf{x})} |\mathbf{x}\rangle$ . The QCBM model loss can then be rewritten as the expectation of an observable, e.g. in the form of Eq. (19), with  $\rho = |\phi\rangle\langle\phi|$  and  $O = |\mathbf{0}\rangle\langle\mathbf{0}|$  being the all-zero projective measurement. Crucially, as  $O = |\mathbf{0}\rangle\langle\mathbf{0}|$  is a global projector, the quantum fidelity is subject to a globality-induced barren plateaus [31] and the loss exponentially concentrates towards one [23]. That is, we have

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{QF}(\boldsymbol{\theta})] \in \mathcal{O}(1/b^n). \quad (47)$$

This global-measurement-induced barren plateau can however be avoided by localising  $\mathcal{L}_{QF}(\boldsymbol{\theta})$ . That is, we replace the global projective measurement  $|\mathbf{0}\rangle\langle\mathbf{0}|$  with its local version  $H_L = \frac{1}{n} \sum_{i=1}^n |0_i\rangle\langle 0_i| \otimes \mathbb{1}_{\bar{i}}$ , where  $\bar{i}$  indicates all qubits except qubit  $i$ . The new localised version of the quantum fidelity loss is given by

$$\mathcal{L}_{QF}^{(L)}(\boldsymbol{\theta}) = 1 - \langle\phi| U(\boldsymbol{\theta}) H_L U^\dagger(\boldsymbol{\theta}) |\phi\rangle. \quad (48)$$

This local loss is faithful to its global variant for product state training in the sense that it vanishes under the same conditions [92], i.e. when the QCBM distribution matches the data distribution exactly. However, it enjoys trainability guarantees via the results of Ref. [31]. This implies that, unlike the MMD and other classical losses that utilize the conventional measurement strategy, the local quantum fidelity can effectively distinguish between high-order marginals even at  $k = 1$  bodyness. However, although the local loss function can evade global measurement-induced BPs, it still suffers under BPs from other sources, such as expressibility or noise. Additionally, it is not yet explored how practical a fidelity loss is for the purposes of generalizing from training data.

Fig. 7 depicts numerical variance results for the fidelity loss on a range of datasets, circuit depths and numbers of

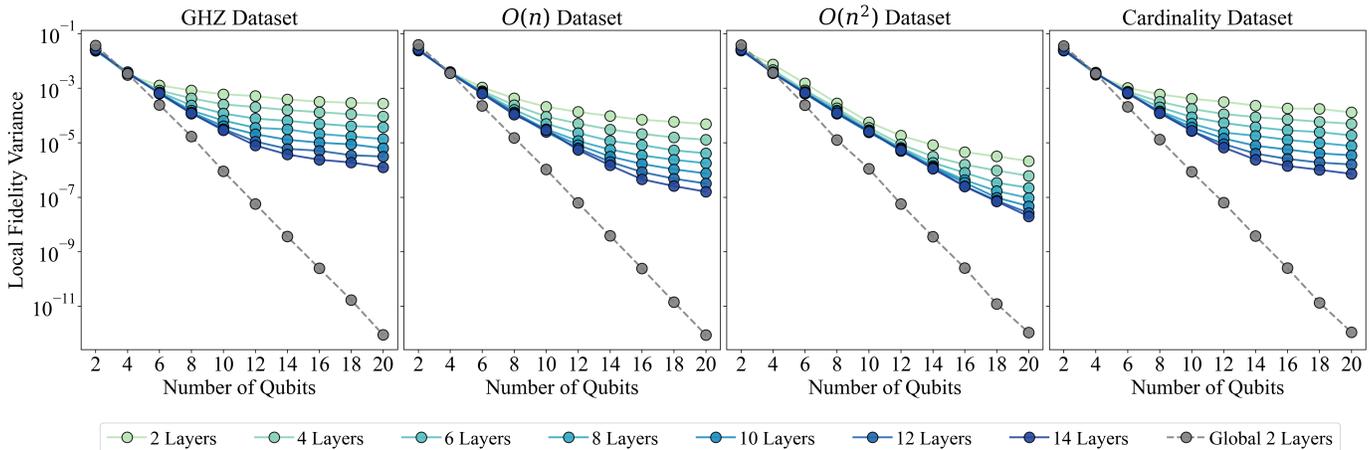


Figure 7. **Study of loss concentration with the local quantum fidelity loss function.** Numerical evidence that the local quantum fidelity loss function does not exhibit global or explicit loss function barren plateaus. It does however exhibit expressivity-induced barren plateaus, as it is the case for all VQA-type loss functions in the form of Eq. (19). In contrast, the global quantum fidelity variance decays exponentially at all circuit depths. The numerical setup is the same as for the MMD in Fig. 6.

qubits. For all datasets, the local quantum fidelity exhibits only polynomially decaying variance over random parameters when the quantum circuits are not too deep. As a reference, we additionally depict the global quantum fidelity which exponentially decays for all circuit depths.

The challenge now becomes how to estimate  $\mathcal{L}_{QF}^{(L)}(\theta)$  using measurements from the quantum computer. The seemingly straight-forward approach is to prepare the initial state  $|\phi\rangle$ , evolve it under  $U^\dagger(\theta)$ , and then evaluate the observable defined by  $H_L$  through measurements in the computational basis. However, loading classical data into a quantum state  $|\phi\rangle$  is not expected to be feasible in general. In Appendix D, we propose an approach that can be used to estimate  $\mathcal{L}_{QF}^{(L)}(\theta)$  using a series of Hadamard tests without needing to prepare  $|\phi\rangle$ . We note that, while in theory our approach requires a number of Hadamard tests that scales with the amount training data, we expect stochastic techniques, such as stochastic gradient descent [93], to be sufficient in practice.

*Broader Implications.* In this section we have presented one example of a quantum strategy to measure a fidelity-based loss for quantum generative modelling. While this approach puts more load on the quantum computer as compared to losses employing the conventional measurement strategy, it enjoys simultaneous trainability and faithfulness to the target distribution.

An interesting extension would be to explore other quantum approaches for efficiently training QML models. One could for example attempt to compute the KL divergence

or other explicit losses directly on the quantum computer. Although the implementation of non-linear operations on quantum computers has been demonstrated in Ref. [94–96], we are not yet aware of quantum strategies beyond one related demonstration for the Rényi divergence in Ref. [97]. One alternative approach would be to attempt to indirectly turn the QCBM into an explicit generative model by estimating its probabilities using amplitude amplification or other techniques.

#### IV. TRAINING ON A HEP DATASET

In this section, we perform realistic training of QCBMs on a more practical dataset which is derived from HEP colliders experiments. We compare the implicit cost functions MMD and local quantum fidelity (LQF) with the explicit KL divergence for an increasing number of circuit depth  $L$  and the number of qubits  $n$ , and across several measurement budgets. To summarize our results, we observe that the presence of shot noise causes the training with KLD to fail, while the MMD and LQF both hold up significantly better.

*Dataset.* We consider a dataset consisting of energy depositions in an electromagnetic calorimeter (ECAL) [98]. The data was generated using a Monte Carlo approach (theGeant4 toolkit [99]), which accurately describes the ECAL detector behaviour under a typically proton-proton collision at a LHC experiment. The dataset consists of the energy deposition on a  $25 \times 25 \times 25$  grid, that we

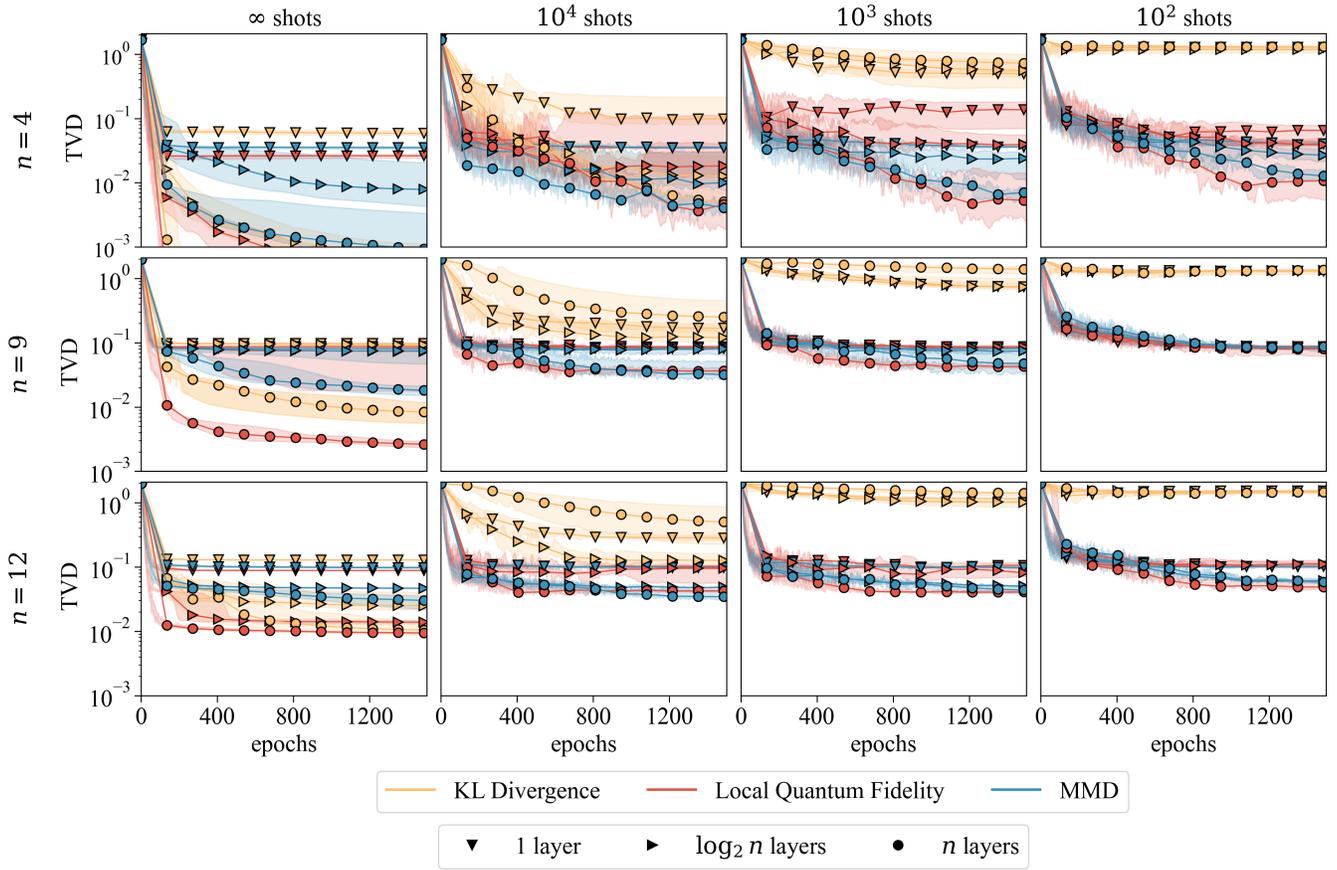


Figure 8. **Finite-shot comparison of loss functions.** TVD, computed with infinite statistics, on the training curve of the QCBMs with varying number of qubits  $n = 4$ ,  $n = 9$  and  $n = 12$  (rows) and layers (symbols), where the gradients are computed with different number of shots (columns) for different loss function (colours).

downsized to a two dimensional grid of various sizes. The images are converted to a black and white scale by considering the pixel ‘hit’ if the energy deposition exceeds a certain threshold, which is chosen as one tenth of the mean energy deposit. We map each pixel to a qubit and take the state  $|1\rangle$  to represent a hit. This dataset naturally has a polynomial support and thus is precisely the type of dataset that we might hope to learn using quantum machine learning.

*Training.* We use a parametrized quantum circuit of the form

$$U(\theta) = \left[ \prod_{l=1}^L R(\theta_l) W_l(\alpha_l) \right] R(\theta_0), \quad (49)$$

where  $R(\theta_l)$  is a layer of arbitrary single qubit unitaries that can be parameterised using  $3n$  Euler angles,  $W(\alpha_l) := \prod_{i=1}^{n-1} CX_{i,i+1} RY_i(\alpha_l^i) CX_{i,i+1}$  acts as parametrized entan-

gling gate with  $CX_{i,j}$  a CNOT gate between qubits  $i$  and  $j$  and  $RY_i(\alpha_l^i)$  a single qubit rotation of qubit  $i$  around the  $y$ -axis, and the parameters  $\theta = \{\theta_l, \alpha_l\}$ . We use the total variation distance (TVD), see Eq. (7), as a common metric to assess the performance of each loss function. To verify performance accurately, we compute the TVD using exact simulation. The gradients for each loss are computed using the parameter shift rule [100] which provides estimates of the analytical gradient, and the parameters are updated with the ADAM [101] optimizer with a decaying learning rate  $\text{lr}(t) = \max(0.01e^{-\beta t}, 10^{-5})$ , where  $t$  is the optimization step and  $\beta = 0.005$  the decaying rate. The computation of the KLD is stabilised using a regulariser of  $\epsilon = 10^{-6}$ , which has been tuned through trial and error. To follow best-practices with the MMD, we average the gradient estimation over several different bandwidths

$$\sigma = (0.01, 0.1, 0.25, 0.5, 1, 10) n, \quad (50)$$

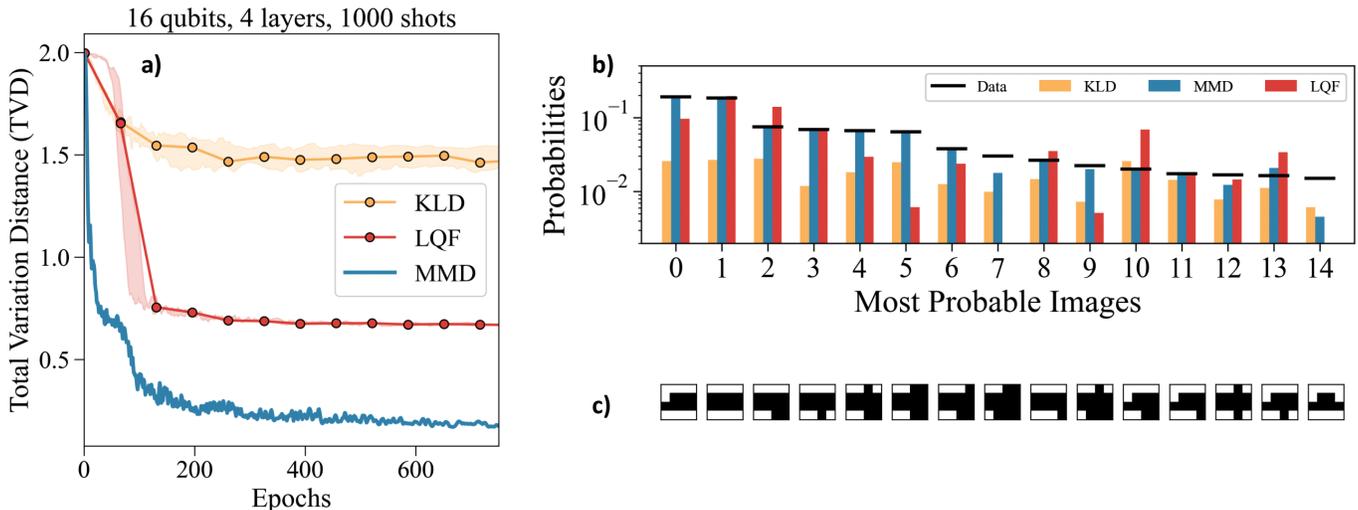


Figure 9. **16 qubit QCBM training.** a) TVD, computed with infinite statistics on the training curve of  $n = 16$  QCBM with  $\log_2 n = 4$  layers trained with 1000 shots. Since the computations are more expensive, only one training curve is shown for the MMD, while for the KLD and LQF, ten and six repetitions are performed respectively. b) Histograms of the trained QCBMs, where only the 15 most-occurring images, shown in panel c), are displayed. The black lines denote the target probabilities.

which incurs no additional quantum resources. This makes the loss full-bodied and thus keeps the model trainable while aiding convergence. We note that these are likely not the optimal bandwidths to average over but it demonstrates the best-practice approach.

*Results.* Fig. 8 shows the TVD, computed with infinite statistics, on the training curve of the QCBMs with varying number of qubits  $n \in \{4, 9, 12\}$  (rows) and layers (symbols), where the gradients are computed with different number of shots (columns) for different loss function (colours). The lines denote the median over ten random parameter initialisation while the shaded area denotes the 25% to 75% percentile. We observe that the performance of the KLD quickly deteriorates as the number of shots is reduced while the MMD and local quantum fidelity (LQF) remain more stable. We further observe that increasing the expressivity of the QCBM from  $\log_2 n$  to  $n$  layers does not lead to a significant increase in performance for a low number of shots.

To demonstrate the scalability to larger systems, we also train a  $n = 16$  QCBM with  $\log_2 n = 4$  layers and 1000 shots per function evaluation. We performed a manual search over different ansatz structures and used the best performing one across the losses. Hence, a slightly different ansatz from Eq. (49) is used, where the difference is in the entanglement map, which is replaced by  $\prod_{i < j} CX_{i,j}$ . In panel a) of Fig. 9, we depict the median and 25% to 75% percentiles for the KLD and LQF over 10 and 6 ran-

dom repetitions respectively, whereas we only show a single representative example for the MMD. In panel b) we show the probability histograms of the 15 most occurring images in the dataset, as well as the final respective model probabilities. The corresponding  $4 \times 4$  pixel images are displayed at the bottom panel c).

In this 16-qubit example, it appears that the LQF is no longer performing on-par with the MMD, as was the case in Fig. 8 for smaller system sizes. A possible explanation is that one chooses all relative phases in the data state  $|\phi\rangle = \sum_{\mathbf{x}} \sqrt{\tilde{p}(\mathbf{x})} |\mathbf{x}\rangle$ , which strongly reduces the number of wavefunctions that minimize the LQF loss even though they may produce the desired measurement distribution. This may not only produce less solutions, it also enforces that the ansatz needs to be able to express exactly that state. While this could be leveraged using specialized real-valued ansätze, this is not attempted here. We conclude that the practical properties of the LQF loss as compared to implicit losses using the conventional measurement strategy are still to be studied in more detail.

To emphasise the importance of the size of the support, in Appendix E we also consider an exponential version of the dataset, by using a negative logarithm transformation. We find in this case that the KLD does not suffer from exponential concentration and can be trained. This explains the successes previously observed for training QCBMs using the KLD for small scale problems. However, as the amount of classical training data cannot scale exponentially

these successes are not relevant to larger, non-classically simulable, problems.

## V. DISCUSSION

In this work, we have introduced the notion of explicit and implicit losses, which broadly reflect the capabilities of explicit and implicit generative models [45]. We argue that these concepts provide a useful framework to understand the trainability of quantum generative models. In particular, we argue that the mismatch between the indirect access to probabilities provided by implicit models with the explicit probabilities required by explicit losses renders implicit models untrainable via explicit losses. More concretely, focusing our attention on quantum circuit Born machines as a commonly used implicit model, we prove that pairwise explicit losses exponentially concentrate (Theorem 1). This result prohibits efficient training using a large class of commonly-used losses including the KL divergence, JS divergence and the total variation distance. Such losses may however be usable with explicit “quantum” generative models such as tensor network Born machines [55, 56, 102].

Crucially, our results assume access to a polynomial (in the number of qubits) number of training data samples and measurements from the quantum circuit. With only moderate numbers of qubits, this assumption is unnecessary and explicit losses such as the KL divergence may appear to be trainable (see e.g. Refs. [16, 17, 19–21]). More generally, if we restrict the number of qubits used to classically simulable sizes, this assumption can be lifted and one could use quantum generative models purely for their efficient sampling capabilities. However, to harness the full potential of quantum generative modelling one surely wants to push to non-classically tractable problem sizes, at which point this assumption is essential. For example, even with only 50 qubits, access to  $\sim 2^{50} \approx 10^{15}$  training samples or quantum measurements is unrealistic.

While formulated initially for random quantum circuits, the intuition underlying Theorem 1 suggests our no-go result extends to scenarios where the implicit generative model’s measurement distribution only has polynomial support (e.g., near-identity initialization of the circuit [103]), as well as beyond the pairwise explicit form of the explicit loss. One exception may be if the quantum generative model has a strong inductive bias. Hence our work further motivates the search for new methods for constructing parameterised circuits with strong inductive biases (e.g. via warm starts [104] or incorporating symmetry constraints [41, 105–108]).

In contrast to explicit losses, implicit losses are naturally suited to training implicit models. Within this line of

thought, we have identified the MMD loss with a Gaussian kernel as a promising implicit loss for training QCBMs. We show that this loss can be interpreted as the expectation value of a quantum observable, where crucially the properties of the observable depend on the bandwidth parameter  $\sigma$ . In the common case where  $\sigma$  is independent of the system size,  $\sigma \in \mathcal{O}(1)$ , the observable becomes predominantly global and thus exponentially concentrates. Conversely, when  $\sigma$  scales linearly with the system size,  $\sigma \in \Theta(n)$ , the low-body interaction terms in the observable are largely dominant over the global terms and hence exhibits large gradients. We provide a rigorous theoretical guarantee for the case of a QCBM with a randomly initialized tensor product ansatz, showing that the MMD variance scales at least polynomially with the number of qubits. Based on the low-body operator interpretation of the MMD loss and numerical evidence up to 20 qubits, we further argue that the MMD loss with  $\sigma \in \Theta(n)$  should remain trainable for quantum circuits with  $\mathcal{O}(\log(n))$  depth.

Our main results for explicit and implicit losses assume the conventional strategy for estimating a generative loss function from an implicit model, where the model provides a set of samples in the computational basis, which are then used to estimate the loss in conjunction with the training data samples. While this is the standard classical strategy, quantum generative models can employ alternative quantum strategies by leveraging quantum computing power. As an example, we propose the local quantum fidelity as a trainable loss function for generative modelling. Developing alternative quantum strategies for training quantum generative models is an interesting avenue for future research. A natural candidate might be, as suggested in Ref. [13], to implement the MMD loss with a *quantum* kernel, where the kernel values themselves are estimated using quantum computers. While one could hope for a potential quantum advantage with this approach (especially when training on quantum data), there is the additional challenge that quantum kernels without inductive bias tend to exponentially concentrate [29].

To put our conclusions to the test, we studied how these loss functions perform in more practical scenarios with data derived from High Energy Physics experiments at the LHC. This dataset naturally satisfies our assumptions of a polynomial number of data samples at all system sizes. Our training results are found to be consistent with our theoretical predictions in which both the MMD and quantum fidelity losses significantly outperform the KLD loss when a strict measurement budget is employed.

Finally, while our work addresses the question of whether a given loss exhibits non-exponentially vanishing gradients, we stress that this is just one ingredient among many to ensure the success of quantum generative modelling. Of

particular importance is the observation that models with local losses will generally struggle to learn global correlations due to the function’s inability to distinguish high-order features in the data. Hence we advocate using *full-body* losses, which contain both low and high-body terms, for quantum generative modelling. More broadly, the ability of a model to successfully generalize will also presumably depend on the choice in loss, but this is beyond the scope of this work. Nonetheless, ensuring non-vanishing loss gradients and ensuring faithfulness of the loss function are critical steps since failing here precludes the successful training and generalization of quantum generative models.

Hence, our work constitutes an important first step to understanding of the barriers that need to be overcome to achieve a quantum advantage in generative modelling.

## ACKNOWLEDGMENTS

The authors would like to acknowledge Christa Zoufal and Marco Cerezo for helpful discussions. ST and ZH acknowledge support from the Sandoz Family Foundation-Monique de Meuron program for Academic Promotion. OK, SV and MG are supported by CERN through the CERN Quantum Technology Initiative.

- 
- [1] Aram W Harrow and Ashley Montanaro, “Quantum computational supremacy,” *Nature* **549**, 203–209 (2017).
  - [2] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, “Quantum machine learning,” *Nature* **549**, 195–202 (2017).
  - [3] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sit-tan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean, “Quantum advantage in learning from experiments,” *Science* **376**, 1182–1186 (2022).
  - [4] Andrew J Daley, Immanuel Bloch, Christian Kokail, Stuart Flannigan, Natalie Pearson, Matthias Troyer, and Peter Zoller, “Practical quantum advantage in quantum simulation,” *Nature* **607**, 667–676 (2022).
  - [5] John Preskill, “Quantum computing in the NISQ era and beyond,” *Quantum* **2**, 79 (2018).
  - [6] Aram W Harrow, Avinandan Hassidim, and Seth Lloyd, “Quantum algorithm for linear systems of equations,” *Physical Review Letters* **103**, 150502 (2009).
  - [7] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost, “Quantum principal component analysis,” *Nature Physics* **10**, 631–633 (2014).
  - [8] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean, “Power of data in quantum machine learning,” *Nature Communications* **12**, 1–9 (2021).
  - [9] Eric R Anschuetz, Hong-Ye Hu, Jin-Long Huang, and Xun Gao, “Interpretable quantum advantage in neural sequence learning,” *arXiv preprint arXiv:2209.14353* (2022).
  - [10] Javier Alcazar, Vicente Leyton-Ortega, and Alejandro Perdomo-Ortiz, “Classical versus quantum models in machine learning: insights from a finance application,” *Machine Learning: Science and Technology* **1**, 035003 (2020).
  - [11] Kaitlin Gili, Mohamed Hibat-Allah, Marta Mauri, Chris Ballance, and Alejandro Perdomo-Ortiz, “Do quantum circuit born machines generalize?” *arXiv preprint arXiv:2207.13645* (2022).
  - [12] Alejandro Perdomo-Ortiz, Marcello Benedetti, John Realpe-Gómez, and Rupak Biswas, “Opportunities and challenges for quantum-assisted machine learning in near-term quantum computers,” *Quantum Science and Technology* **3**, 030502 (2018).
  - [13] Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi, “The born supremacy: quantum advantage and training of an ising born machine,” *npj Quantum Information* **6** (2020).
  - [14] Ryan Sweke, Jean-Pierre Seifert, Dominik Hangleiter, and Jens Eisert, “On the quantum versus classical learnability of discrete distributions,” *Quantum* **5**, 417 (2021).
  - [15] Xun Gao, Eric R. Anschuetz, Sheng-Tao Wang, J. Ignacio Cirac, and Mikhail D. Lukin, “Enhancing generative models via quantum correlations,” *Phys. Rev. X* **12**, 021037 (2022).
  - [16] Manuel S Rudolph, Ntwali Bashige Toussaint, Amara Katabarwa, Sonika Johri, Borja Peropadre, and Alejandro Perdomo-Ortiz, “Generation of high-resolution handwritten digits with an ion-trap quantum computer,” *Physical Review X* **12**, 031010 (2022).
  - [17] Brian Coyle, Maxwell Henderson, Justin Chan Jin Le, Niraj Kumar, Marco Paini, and Elham Kashefi, “Quantum versus classical generative modelling in finance,” *Quantum Science and Technology* **6**, 024013 (2021).
  - [18] Andrea Delgado and Kathleen E. Hamilton, “Unsupervised quantum circuit learning in high energy physics,” *Phys. Rev. D* **106**, 096006 (2022).
  - [19] Kathleen E Hamilton, Eugene F Dumitrescu, and Raphael C Pooser, “Generative model benchmarks for superconducting qubits,” *Physical Review A* **99**, 062323 (2019).
  - [20] Vicente Leyton-Ortega, Alejandro Perdomo-Ortiz, and Oscar Perdomo, “Robust implementation of generative modeling with parametrized quantum circuits,” *Quantum Machine Intelligence* **3**, 1–10 (2021).
  - [21] Daiwei Zhu, Norbert M Linke, Marcello Benedetti, Kevin A Landsman, Nhung H Nguyen, C Huerta Alderete,

- Alejandro Perdomo-Ortiz, Nathan Korda, A Garfoot, Charles Brecque, *et al.*, “Training of quantum circuits on a hybrid quantum computer,” *Science advances* **5** (2019), [10.1126/sciadv.aaw9918](https://doi.org/10.1126/sciadv.aaw9918).
- [22] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven, “Barren plateaus in quantum neural network training landscapes,” *Nature Communications* **9**, 1–6 (2018).
- [23] Andrew Arrasmith, Zoë Holmes, Marco Cerezo, and Patrick J Coles, “Equivalence of quantum barren plateaus to cost concentration and narrow gorges,” *Quantum Science and Technology* **7**, 045015 (2022).
- [24] Martín Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo, “Diagnosing Barren Plateaus with Tools from Quantum Optimal Control,” *Quantum* **6**, 824 (2022).
- [25] M. Cerezo and Patrick J Coles, “Higher order derivatives of quantum neural networks with barren plateaus,” *Quantum Science and Technology* **6**, 035006 (2021).
- [26] Andrew Arrasmith, M. Cerezo, Piotr Czarnik, Lukasz Cincio, and Patrick J Coles, “Effect of barren plateaus on gradient-free optimization,” *Quantum* **5**, 558 (2021).
- [27] 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).
- [28] Chen Zhao and Xiao-Shan Gao, “Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus,” *Quantum* **5**, 466 (2021).
- [29] Supanut Thanasilp, Samson Wang, M. Cerezo, and Zoë Holmes, “Exponential concentration and untrainability in quantum kernel methods,” *arXiv preprint arXiv:2208.11060* (2022).
- [30] Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J Coles, “Connecting ansatz expressibility to gradient magnitudes and barren plateaus,” *PRX Quantum* **3**, 010313 (2022).
- [31] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles, “Cost function dependent barren plateaus in shallow parametrized quantum circuits,” *Nature Communications* **12**, 1–12 (2021).
- [32] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe, “Entanglement-induced barren plateaus,” *PRX Quantum* **2**, 040316 (2021).
- [33] Taylor L Patti, Khadijeh Najafi, Xun Gao, and Susanne F Yelin, “Entanglement devised barren plateau mitigation,” *Physical Review Research* **3**, 033090 (2021).
- [34] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles, “Noise-induced barren plateaus in variational quantum algorithms,” *Nature Communications* **12**, 1–11 (2021).
- [35] Samson Wang, Piotr Czarnik, Andrew Arrasmith, M. Cerezo, Lukasz Cincio, and Patrick J Coles, “Can error mitigation improve trainability of noisy variational quantum algorithms?” *arXiv preprint arXiv:2109.01051* (2021).
- [36] Supanut Thanasilp, Samson Wang, Nhat A Nghiem, Patrick J. Coles, and M. Cerezo, “Subtleties in the trainability of quantum machine learning models,” *arXiv preprint arXiv:2110.14753* (2021).
- [37] Lorenzo Leone, Salvatore FE Oliviero, Lukasz Cincio, and M Cerezo, “On the practical usefulness of the hardware efficient ansatz,” *arXiv preprint arXiv:2211.01477* (2022).
- [38] Guangxi Li, Ruilin Ye, Xuanqiang Zhao, and Xin Wang, “Concentration of data encoding in parameterized quantum circuits,” *arXiv preprint arXiv:2206.08273* (2022).
- [39] John Napp, “Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze,” *arXiv preprint arXiv:2203.06174* (2022).
- [40] Arthur Pesah, M. Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles, “Absence of barren plateaus in quantum convolutional neural networks,” *Physical Review X* **11**, 041011 (2021).
- [41] Martín Larocca, Frédéric Sauvage, Faris M. Sbahi, Guillaume Verdon, Patrick J. Coles, and M. Cerezo, “Group-invariant quantum machine learning,” *PRX Quantum* **3**, 030341 (2022).
- [42] Jirawat Tangpanitanon, Supanut Thanasilp, Ninnat Dangniam, Marc-Antoine Lemonde, and Dimitris G Angelakis, “Expressibility and trainability of parametrized analog quantum systems for machine learning applications,” *Physical Review Research* **2**, 043364 (2020).
- [43] Kunal Sharma, M. Cerezo, Lukasz Cincio, and Patrick J Coles, “Trainability of dissipative perceptron-based quantum neural networks,” *Physical Review Letters* **128**, 180505 (2022).
- [44] Manuel S Rudolph, Sukin Sim, Asad Raza, Michal Stechly, Jarrod R McClean, Eric R Anschuetz, Luis Serano, and Alejandro Perdomo-Ortiz, “Orqviz: Visualizing high-dimensional landscapes in variational quantum algorithms,” *arXiv preprint arXiv:2111.04695* (2021).
- [45] Shakir Mohamed and Balaji Lakshminarayanan, “Learning in implicit generative models,” *arXiv preprint arXiv:1610.03483* (2016).
- [46] Marcello Benedetti, Delfina Garcia-Pintos, Oscar Perdomo, Vicente Leyton-Ortega, Yunseong Nam, and Alejandro Perdomo-Ortiz, “A generative modeling approach for benchmarking and training shallow quantum circuits,” *npj Quantum Information* **5**, 1–9 (2019).
- [47] I. Csiszar, “ $I$ -Divergence Geometry of Probability Distributions and Minimization Problems,” *The Annals of Probability* **3**, 146 – 158 (1975).
- [48] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola, “A kernel two-sample test,” *Journal of Machine Learning Research* **13**, 723–773 (2012).
- [49] Kaitlin Gili, Marta Mauri, and Alejandro Perdomo-Ortiz, “Evaluating generalization in classical and quantum generative models,” *arXiv preprint arXiv:2201.08770* (2022).
- [50] Song Cheng, Jing Chen, and Lei Wang, “Information perspective to probabilistic modeling: Boltzmann machines versus born machines,” *Entropy* **20**, 583 (2018).
- [51] Jin-Guo Liu and Lei Wang, “Differentiable learning of quantum circuit born machines,” *Phys. Rev. A* **98**, 062324

- (2018).
- [52] P. Smolensky, “Information processing in dynamical systems: Foundations of harmony theory,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (MIT Press, Cambridge, MA, USA, 1986) p. 194–281.
- [53] Geoffrey E Hinton, “A practical guide to training restricted boltzmann machines,” *Neural Networks: Tricks of the Trade: Second Edition*, 599–619 (2012).
- [54] Zhao-Yu Han, Jun Wang, Heng Fan, Lei Wang, and Pan Zhang, “Unsupervised generative modeling using matrix product states,” *Phys. Rev. X* **8**, 031012 (2018).
- [55] Song Cheng, Lei Wang, Tao Xiang, and Pan Zhang, “Tree tensor networks for generative modeling,” *Phys. Rev. B* **99**, 155131 (2019).
- [56] Tom Vieijra, Laurens Vanderstraeten, and Frank Verstraete, “Generative modeling with projected entangled-pair states,” *arXiv preprint arXiv:2202.08177* (2022).
- [57] Michael L. Wall, Matthew R. Abernathy, and Gregory Quiroz, “Generative machine learning with tensor networks: Benchmarks on near-term quantum computers,” *Phys. Rev. Research* **3**, 023010 (2021).
- [58] Ieva Čepaitė, Brian Coyle, and Elham Kashefi, “A continuous variable born machine,” *Quantum Mach. Intell.* **4**, 6 (2022).
- [59] Oriël Kiss, Michele Grossi, Enrique Kajomovitz, and Sofia Vallecorsa, “Conditional born machine for monte carlo event generation,” *Phys. Rev. A* **106**, 022612 (2022).
- [60] Marcello Benedetti, Brian Coyle, Mattia Fiorentini, Michael Lubasch, and Matthias Rosenkranz, “Variational inference with a quantum computer,” *Phys. Rev. Appl.* **16**, 044057 (2021).
- [61] Kaitlin Gili, Mykolas Sveistrys, and Chris Ballance, “Introducing nonlinear activations into quantum generative models,” *Phys. Rev. A* **107**, 012406 (2023).
- [62] Sofiene Jerbi, Lukas J Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko, “Quantum machine learning beyond kernel methods,” *Nature Communications* **14**, 517 (2023).
- [63] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, “Pixel recurrent neural networks,” in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (2016) p. 1747–1756.
- [64] David E. Rumelhart and James L. McClelland, “Learning internal representations by error propagation,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (1987) pp. 318–362.
- [65] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, Vol. 27 (2014).
- [66] Imre Csiszár, “On information-type measure of difference of probability distributions and indirect observations,” *Studia Sci. Math. Hungar.* **2**, 299–318 (1967).
- [67] Solomon Kullback and Richard A Leibler, “On information and sufficiency,” *The annals of mathematical statistics* **22**, 79–86 (1951).
- [68] J. Lin, “Divergence measures based on the shannon entropy,” *IEEE Transactions on Information Theory* **37**, 145–151 (1991).
- [69] Alfréd Rényi, “On measures of entropy and information,” in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (1961) pp. 547–562.
- [70] Joe Gibbs, Kaitlin Gili, Zoë Holmes, Benjamin Commeau, Andrew Arrasmith, Lukasz Cincio, Patrick J. Coles, and Andrew Sornborger, “Long-time simulations with high fidelity on quantum hardware,” *arXiv preprint arXiv:2102.04313* (2021).
- [71] Joe Gibbs, Zoe Holmes, Matthias C. Caro, Nicholas Ezzell, Hsin-Yuan Huang, Lukasz Cincio, Andrew T. Sornborger, and Patrick J. Coles, “Dynamical simulation via quantum machine learning with provable generalization,” *arXiv preprint arXiv:2204.10269* (2022).
- [72] Matthias C. Caro, Hsin-Yuan Huang, Nicholas Ezzell, Joe Gibbs, Andrew T. Sornborger, Lukasz Cincio, Patrick J. Coles, and Zoe Holmes, “Out-of-distribution generalization for learning quantum dynamics,” *arXiv preprint arXiv:2204.10268* (2022).
- [73] Tyler Volkoff, Zoë Holmes, and Andrew Sornborger, “Universal compiling and (no-)free-lunch theorems for continuous-variable quantum learning,” *PRX Quantum* **2**, 040327 (2021).
- [74] Adriano Barenco, Andre Berthiaume, David Deutsch, Artur Ekert, Richard Jozsa, and Chiara Macchiavello, “Stabilization of quantum computations by symmetrization,” *SIAM Journal on Computing* **26**, 1541–1557 (1997).
- [75] Juan Carlos Garcia-Escartin and Pedro Chamorro-Posada, “Swap test and hong-ou-mandel effect are equivalent,” *Physical Review A* **87**, 052330 (2013).
- [76] Seth Lloyd and Christian Weedbrook, “Quantum generative adversarial learning,” *Phys. Rev. Lett.* **121**, 040502 (2018).
- [77] Christa Zoufal, Aurélien Lucchi, and Stefan Woerner, “Quantum generative adversarial networks for learning and loading random distributions,” *npj Quantum Information* **5**, 103 (2019).
- [78] Haozhen Situ, Zhimin He, Yuyi Wang, Lvzhou Li, and Shenggen Zheng, “Quantum generative adversarial network for generating discrete distribution,” *Information Sciences* **538**, 193–208 (2020).
- [79] Carlos Bravo-Prieto, Julien Baglio, Marco Cè, Anthony Francis, Dorota M. Grabowska, and Stefano Carrazza, “Style-based quantum generative adversarial networks for Monte Carlo events,” *Quantum* **6**, 777 (2022).
- [80] Murphy Yuezhen Niu, Alexander Zlokapa, Michael Broughton, Sergio Boixo, Masoud Mohseni, Vadim Smelyanskiy, and Hartmut Neven, “Entangling quantum generative adversarial networks,” *Phys. Rev. Lett.* **128**, 220505 (2022).
- [81] Daniel Stilck França and Raul Garcia-Patron, “Limitations of optimization algorithms on noisy quantum de-

- vices,” *Nature Physics* **17**, 1221–1227 (2021).
- [82] T-J Chang, Nigel Meade, John E Beasley, and Yazid M Sharaiha, “Heuristics for cardinality constrained portfolio optimisation,” *Computers & Operations Research* **27**, 1271–1302 (2000).
- [83] Javier Alcazar, Mohammad Ghazi Vakili, Can B Kalayci, and Alejandro Perdomo-Ortiz, “Geo: Enhancing combinatorial optimization with classical and quantum generative models,” *arXiv preprint arXiv:2101.06250* (2021).
- [84] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos, “Mmd gan: Towards deeper understanding of moment matching network,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017) p. 2200–2210.
- [85] Wei Wang, Yuan Sun, and Saman Halgamuge, “Improving mmd-gan training with repulsive loss function,” *arXiv preprint arXiv:1812.09916* (2018).
- [86] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot, “Domain generalization with adversarial feature learning,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018) pp. 5400–5409.
- [87] Christian Igel, Nikolaus Hansen, and Stefan Roth, “Covariance Matrix Adaptation for Multi-objective Optimization,” *Evolutionary Computation* **15**, 1–28 (2007).
- [88] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur, “Optimal kernel choice for large-scale two-sample tests,” in *Advances in Neural Information Processing Systems* (2012).
- [89] Danica J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton, “Generative models and model criticism via optimized maximum mean discrepancy,” *arXiv preprint arXiv:1611.04488* (2016).
- [90] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa, “Large sample analysis of the median heuristic,” *arXiv preprint arXiv:1707.07269* (2017).
- [91] Chiara Leadbeater, Louis Sharrock, Brian Coyle, and Marcello Benedetti, “F-divergences and cost function locality in generative modelling with quantum circuits,” *Entropy* **23**, 1281 (2021).
- [92] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles, “Quantum-assisted quantum compiling,” *Quantum* **3**, 140 (2019).
- [93] 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).
- [94] Zoë Holmes, Nolan J Coble, Andrew T Sornborger, and Yiğit Subaşı, “Nonlinear transformations in quantum computation,” *Physical Review Research* **5**, 013105 (2023).
- [95] Andrés Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe, “Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics,” *arXiv preprint arXiv:1806.01838* (2018).
- [96] John M. Martyn, Zane M. Rossi, Andrew K. Tan, and Isaac L. Chuang, “Grand unification of quantum algorithms,” *PRX Quantum* **2**, 040203 (2021).
- [97] Maria Kieferova, Ortiz Marrero Carlos, and Nathan Wiebe, “Quantum generative training using rényi divergences,” *arXiv preprint arXiv:2106.09567* (2021).
- [98] Maurizio Pierini and Matt Zhang, “CLIC Calorimeter 3D images: Electron showers at Fixed Angle,” (2020), [10.5281/zenodo.3603122](https://doi.org/10.5281/zenodo.3603122).
- [99] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, F. Behner, L. Bellagamba, J. Boudreau, L. Broglia, A. Brunengo, H. Burkhardt, S. Chauvie, J. Chuma, R. Chytracek, G. Cooperman, G. Cosmo, P. Degtyarenko, A. Dell’Acqua, G. Depaola, D. Dietrich, R. Enami, A. Feliciello, C. Ferguson, H. Fesefeldt, G. Folger, F. Foppiano, A. Forti, S. Garelli, S. Giani, R. Giannitrapani, D. Gibin, J.J. Gómez Cadenas, I. González, G. Gracia Abril, G. Greeniaus, W. Greiner, V. Grichine, A. Grossheim, S. Guatelli, P. Gumplinger, R. Hamatsu, K. Hashimoto, H. Hasui, A. Heikkinen, A. Howard, V. Ivanchenko, A. Johnson, F.W. Jones, J. Kallenbach, N. Kanaya, M. Kawabata, Y. Kawabata, M. Kawaguti, S. Kelner, P. Kent, A. Kimura, T. Kodama, R. Kokoulin, M. Kossov, H. Kurashige, E. Lamanna, T. Lampén, V. Lara, V. Lefebure, F. Lei, M. Liendl, W. Lockman, F. Longo, S. Magni, M. Maire, E. Medernach, K. Minamimoto, P. Mora de Freitas, Y. Morita, K. Murakami, M. Nagamatu, R. Nartallo, P. Nieminen, T. Nishimura, K. Ohtsubo, M. Okamura, S. O’Neale, Y. Oohata, K. Paech, J. Perl, A. Pfeiffer, M.G. Pia, F. Ranjard, A. Rybin, S. Sadilov, E. Di Salvo, G. Santin, T. Sasaki, N. Savvas, Y. Sawada, S. Scherer, S. Sei, V. Sirotenko, D. Smith, N. Starkov, H. Stoecker, J. Sulkimo, M. Takahata, S. Tanaka, E. Tcherniaev, E. Safai Tehrani, M. Tropeano, P. Truscott, H. Uno, L. Urban, P. Urban, M. Verderi, A. Walkden, W. Wander, H. Weber, J.P. Wellisch, T. Wenaus, D.C. Williams, D. Wright, T. Yamada, H. Yoshida, and D. Zschesche, “Geant4—a simulation toolkit,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506**, 250–303 (2003).
- [100] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran, “Evaluating analytic gradients on quantum hardware,” *Phys. Rev. A* **99**, 032331 (2019).
- [101] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).
- [102] Zhao-Yu Han, Jun Wang, Heng Fan, Lei Wang, and Pan Zhang, “Unsupervised generative modeling using matrix product states,” *Phys. Rev. X* **8**, 031012 (2018).

- [103] Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti, “An initialization strategy for addressing barren plateaus in parametrized quantum circuits,” *Quantum* **3**, 214 (2019).
- [104] Manuel S Rudolph, Jacob Miller, Jing Chen, Atithi Acharya, and Alejandro Perdomo-Ortiz, “Synergy between quantum circuits and tensor networks: Short-cutting the race to practical quantum advantage,” *arXiv preprint arXiv:2208.13673* (2022).
- [105] Louis Schatzki, Martin Larocca, Frederic Sauvage, and M. Cerezo, “Theoretical guarantees for permutation-equivariant quantum neural networks,” *arXiv preprint arXiv:2210.09974* (2022).
- [106] Quynh T. Nguyen, Louis Schatzki, Paolo Braccia, Michael Ragone, Martin Larocca, Frederic Sauvage, Patrick J. Coles, and M. Cerezo, “A theory for equivariant quantum neural networks,” *arXiv preprint arXiv:2210.08566* (2022).
- [107] Michael Ragone, Quynh T. Nguyen, Louis Schatzki, Paolo Braccia, Martin Larocca, Frederic Sauvage, Patrick J. Coles, and M. Cerezo, “Representation theory for geometric quantum machine learning,” *arXiv preprint arXiv:2210.07980* (2022).
- [108] Johannes Jakob Meyer, Marian Mularski, Elies Gil-Fuster, Antonio Anna Mele, Francesco Arzani, Alissa Wilms, and Jens Eisert, “Exploiting symmetry in variational quantum machine learning,” *arXiv preprint arXiv:2205.06217* (2022).
- [109] Daniel A Roberts and Beni Yoshida, “Chaos and complexity by design,” *Journal of High Energy Physics* **2017**, 121 (2017).
- [110] Mohammad H. Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyi, and Roger Melko, “Quantum boltzmann machine,” *Phys. Rev. X* **8**, 021050 (2018).
- [111] 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).

## Appendix

### Appendix A: 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

$$\mathcal{L}_{\text{expl}}(\boldsymbol{\theta}) := \sum_{\mathbf{x}_1 \dots \mathbf{x}_r} f\left(p(\mathbf{x}_1), \dots, p(\mathbf{x}_r), q_{\boldsymbol{\theta}}(\mathbf{x}_1), \dots, q_{\boldsymbol{\theta}}(\mathbf{x}_r)\right), \quad (\text{A1})$$

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

$$\mathcal{L}_{\text{impl}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_r \sim \{p, q_{\boldsymbol{\theta}}\}} g(\mathbf{x}_1, \dots, \mathbf{x}_r), \quad (\text{A2})$$

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. [IIB](#)), 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

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{y}) g(\cdot) \quad (\text{A3})$$

$$= \mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\theta}}, \mathbf{y} \sim p} [g(\cdot)] \quad (\text{A4})$$

where the function  $g$  is for now left arbitrary. By comparing Eq. [\(A3\)](#) and Eq. [\(A1\)](#), 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. [\(A4\)](#) and Eq. [\(A2\)](#), 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. [\(A4\)](#) 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. [\(A1\)](#).

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

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) \delta_{\mathbf{x}, \mathbf{y}} - 2 \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{y}) \delta_{\mathbf{x}, \mathbf{y}} + \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} p(\mathbf{x}) p(\mathbf{y}) \delta_{\mathbf{x}, \mathbf{y}} \\ &= \sum_{\mathbf{x} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})^2 - 2q_{\boldsymbol{\theta}}(\mathbf{x}) p(\mathbf{x}) + p(\mathbf{x})^2 \\ &= \sum_{\mathbf{x} \in \mathcal{X}} (q_{\boldsymbol{\theta}}(\mathbf{x}) - p(\mathbf{x}))^2, \end{aligned} \quad (\text{A5})$$

which takes the form of a pairwise explicit loss (see Eq. [\(3\)](#)). 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. [\(7\)](#) 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. [\(A5\)](#), 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  $\delta$ -trick.

## Appendix B: 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 Appendix B1. In Appendix B2, 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 2 in Appendix B3, 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

$$\mathcal{L}(\theta) = \sum_{\mathbf{x}} f(p(\mathbf{x}), q_{\theta}(\mathbf{x})) , \quad (\text{B1})$$

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. IIB 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

$$\tilde{\mathcal{L}}(\theta) = \sum_{\mathbf{x}} f(\tilde{p}(\mathbf{x}), \tilde{q}_{\theta}(\mathbf{x})) . \quad (\text{B2})$$

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.,  $M = |\tilde{P}|, N = |\tilde{Q}_{\theta}| \in \mathcal{O}(\text{poly}(n))$ .

### 1. 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). 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*

$$\mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta}) = \sum_{\mathbf{x} \in \tilde{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \tilde{Q}_{\theta}} f(0, \tilde{q}_{\theta}(\mathbf{x})) , \quad (\text{B3})$$

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.,

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta})] \geq 1 - \delta , \quad (\text{B4})$$

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  $\Pr \left[ \sum_{\mathbf{x} \in \mathcal{P}} \tilde{q}_\theta(\mathbf{x}) = 0 \right]$  as  $\tilde{q}_\theta(\mathbf{x}) \geq 0$  for all  $\mathbf{x} \in \mathcal{X}$ . So, we have

$$\Pr_{\tilde{Q}_\theta, \theta} [\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)] = \Pr_{\tilde{Q}_\theta, \theta} \left[ \sum_{\mathbf{x} \in \mathcal{P}} \tilde{q}_\theta(\mathbf{x}) = 0 \right] \quad (\text{B5})$$

$$= \int_0^1 \Pr_{\tilde{Q}_\theta} \left[ \sum_{\mathbf{x} \in \mathcal{P}} \tilde{q}_\theta(\mathbf{x}) = 0 \mid \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] \Pr_\theta \left[ \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (\text{B6})$$

$$= \int_0^1 (1-s)^N \Pr_\theta \left[ \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (\text{B7})$$

$$\geq \int_{\mu_s - \sqrt{\sigma_s}}^{\mu_s + \sqrt{\sigma_s}} (1-s)^N \Pr_\theta \left[ \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (\text{B8})$$

$$\geq (1 - (\mu_s + \sqrt{\sigma_s}))^N \int_{\mu_s - \sqrt{\sigma_s}}^{\mu_s + \sqrt{\sigma_s}} \Pr_\theta \left[ \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) = s \right] ds \quad (\text{B9})$$

$$\geq (1 - (\mu_s + \sqrt{\sigma_s}))^N (1 - \sigma_s). \quad (\text{B10})$$

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  $\mu_s$  and  $\sigma_s^2$  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  $\sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x})$

$$\Pr_\theta \left[ \left| \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) - \mu_s \right| \geq k\sigma_s \right] \leq \frac{1}{k^2}. \quad (\text{B11})$$

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

$$\Pr_\theta \left[ \left| \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) - \mu_s \right| \leq \sqrt{\sigma_s} \right] \geq 1 - \sigma_s. \quad (\text{B12})$$

With  $\int_a^b \Pr[x]dx = \Pr[a \leq x \leq b]$ , we then get Eq. (B10).

To further bound the probability, we show that  $\mu_s$  and  $\sigma_s^2$  are exponentially small in the number of qubits. First, consider the mean

$$\mu_s = \mathbb{E}_\theta \left[ \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) \right] \quad (\text{B13})$$

$$= \sum_{\mathbf{x} \in \mathcal{P}} \mu(\mathbf{x}) \quad (\text{B14})$$

$$\leq N_p \max_{\mathbf{x} \in \mathcal{P}} [\mu(\mathbf{x})], \quad (\text{B15})$$

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  $\theta$ . 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

$$\sigma_s^2 = \text{Var}_\theta \left[ \sum_{\mathbf{x} \in \mathcal{P}} q_\theta(\mathbf{x}) \right] \quad (\text{B16})$$

$$= \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_\theta [q_\theta(\mathbf{x})] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \in \mathcal{P} \\ \mathbf{x} \neq \mathbf{x}'}} \text{Cov}_\theta [q_\theta(\mathbf{x}), q_\theta(\mathbf{x}')] \quad (\text{B17})$$

$$\leq \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_\theta [q_\theta(\mathbf{x})] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \in \mathcal{P} \\ \mathbf{x} \neq \mathbf{x}'}} \sqrt{\text{Var}_\theta [q_\theta(\mathbf{x})] \text{Var}_\theta [q_\theta(\mathbf{x}')] } \quad (\text{B18})$$

$$\leq \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_\theta [q_\theta(\mathbf{x})] + \sum_{\substack{\mathbf{x}, \mathbf{x}' \in \mathcal{P} \\ \mathbf{x} \neq \mathbf{x}'}} \frac{\text{Var}_\theta [q_\theta(\mathbf{x})] + \text{Var}_\theta [q_\theta(\mathbf{x}')] }{2} \quad (\text{B19})$$

$$= N_p \sum_{\mathbf{x} \in \mathcal{P}} \text{Var}_\theta [q_\theta(\mathbf{x})] \quad (\text{B20})$$

$$\leq N_p^2 \max_{\mathbf{x} \in \mathcal{P}} [\sigma^2(\mathbf{x})], \quad (\text{B21})$$

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  $q_\theta(\mathbf{x})$  over  $\theta$ . Finally, assuming that  $q_\theta(\mathbf{x}')$  exponentially concentrates over all bitstrings in the dataset, we have  $\sigma_s \in \mathcal{O}(\text{poly}(n)/b^n)$  for some  $b' > 1$ .

Now, we are ready to continue from Eq. (B10).

$$\Pr_{\tilde{Q}_\theta, \theta} [\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_\theta)] \geq (1 - (\mu_s + \sqrt{\sigma_s}))^N (1 - \sigma_s) \quad (\text{B22})$$

$$\geq (1 - N(\mu_s + \sqrt{\sigma_s}))(1 - \sigma_s) \quad (\text{B23})$$

$$\geq 1 - (N(\mu_s + \sqrt{\sigma_s}) + \sigma_s) \quad (\text{B24})$$

$$\geq 1 - \left( N \left( N_p \max_{\mathbf{x} \in \mathcal{P}} [\mu(\mathbf{x})] + \sqrt{N_p \max_{\mathbf{x} \in \mathcal{P}} [\sigma(\mathbf{x})]} \right) + N_p \max_{\mathbf{x} \in \mathcal{P}} [\sigma(\mathbf{x})] \right) \quad (\text{B25})$$

$$= 1 - \delta, \quad (\text{B26})$$

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. (B15) and Eq. (B21). 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

$$\delta \in \mathcal{O} \left( \frac{\text{poly}(n)}{c^n} \right), \quad (\text{B27})$$

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

## 2. 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 1. 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. 2). 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*

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta})] \geq 1 - \delta', \quad \delta' \in \mathcal{O}\left(\frac{\text{poly}(n)}{c^n}\right), \quad (\text{B28})$$

for some  $c' > 1$ , and  $\mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta})$  the statistical estimate of the loss as defined Eq. (20).

*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

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta})] = \frac{\binom{2^n - N_{\mathcal{P}}}{N_{\mathcal{Q}}}}{\binom{2^n}{N_{\mathcal{Q}}}} \quad (\text{B29})$$

$$= \frac{(2^n - N_{\mathcal{P}}) \times \dots \times (2^n - N_{\mathcal{P}} - N_{\mathcal{Q}} + 1)}{2^n \times \dots \times (2^n - N_{\mathcal{P}} + 1)} \quad (\text{B30})$$

$$= \prod_{k=0}^{N_{\mathcal{P}}-1} \frac{2^n - N_{\mathcal{Q}} - k}{2^n - k} \quad (\text{B31})$$

$$= \prod_{k=0}^{N_{\mathcal{P}}-1} \left(1 - \frac{N_{\mathcal{Q}}}{2^n - k}\right) \quad (\text{B32})$$

$$\geq \left(1 - \frac{N_{\mathcal{Q}}}{2^n - N_{\mathcal{P}} + 1}\right)^{N_{\mathcal{P}}} \quad (\text{B33})$$

$$\geq 1 - \frac{N_{\mathcal{Q}}N_{\mathcal{P}}}{2^n - N_{\mathcal{P}} + 1}, \quad (\text{B34})$$

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  $\theta$ . While in the context of VQA-type problems it has been shown to mitigate vanishing gradients at the initial training step [103], 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.

### 3. Untrainability of the pairwise explicit loss function: Proof of Corollary 2

In this sub-section, we provide the proof of Corollary 2, which shows that the concentration of statistical estimate of the pairwise explicit loss makes such losses untrainable.

**Corollary 2** (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  $\theta_1$  and  $\theta_2$  does not contain any information about the training distribution is exponentially close to 1. Particularly, we have*

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) - \tilde{\mathcal{L}}(\theta_2) = \Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2})] \geq 1 - 2\delta, \quad (\text{B35})$$

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  $\theta_1$  (and  $\theta_2$ ), as well as

$$\Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2}) = \sum_{\mathbf{x} \in \mathcal{Q}_{\theta_1}} f(0, \tilde{q}_{\theta_1}(\mathbf{x})) - \sum_{\mathbf{x} \in \mathcal{Q}_{\theta_2}} f(0, \tilde{q}_{\theta_2}(\mathbf{x})), \quad (\text{B36})$$

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

$$\mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_1}) - \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_2}) = \left( \sum_{\mathbf{x} \in \tilde{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \tilde{Q}_{\theta_1}} f(0, \tilde{q}_{\theta_1}(\mathbf{x})) \right) - \left( \sum_{\mathbf{x} \in \tilde{P}} f(\tilde{p}(\mathbf{x}), 0) + \sum_{\mathbf{x} \in \tilde{Q}_{\theta_2}} f(0, \tilde{q}_{\theta_2}(\mathbf{x})) \right) \quad (\text{B37})$$

$$= \sum_{\mathbf{x} \in \tilde{Q}_{\theta_1}} f(0, \tilde{q}_{\theta_1}(\mathbf{x})) - \sum_{\mathbf{x} \in \tilde{Q}_{\theta_2}} f(0, \tilde{q}_{\theta_2}(\mathbf{x})) \quad (\text{B38})$$

$$= \Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2}). \quad (\text{B39})$$

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

$$\Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) - \tilde{\mathcal{L}}(\theta_2) = \Delta\mathcal{L}_0(\tilde{Q}_{\theta_1}, \tilde{Q}_{\theta_2})] = \Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_1}) \cap \tilde{\mathcal{L}}(\theta_2) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_2})] \quad (\text{B40})$$

$$= \Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_1) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_1})] \cdot \Pr_{\tilde{Q}_{\theta}, \theta}[\tilde{\mathcal{L}}(\theta_2) = \mathcal{L}_0(\tilde{P}, \tilde{Q}_{\theta_2})] \quad (\text{B41})$$

$$\geq (1 - \delta)(1 - \delta) \quad (\text{B42})$$

$$\geq 1 - 2\delta, \quad (\text{B43})$$

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  $\delta^2$  term.  $\square$

### Appendix C: 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, Appendix C1 shows how the MMD loss can be viewed as the

expectation of an observable and analyzes how its properties depend on the bandwidth  $\sigma$ . A detailed analysis of the MMD loss landscape for a tensor product QCBM is provided in Appendix C2. Lastly, in Appendix C3, 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

$$\mathcal{L}_{\text{MMD}}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})K(\mathbf{x}, \mathbf{y}) - 2 \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}) + \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} p(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}), \quad (\text{C1})$$

with the classical Gaussian kernel

$$K_{\sigma}(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma}} \quad (\text{C2})$$

$$= \prod_{i=1}^n e^{-\frac{(x_i-y_i)^2}{2\sigma}}, \quad (\text{C3})$$

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.

### 1. 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

$$\mathcal{M}(\rho, \rho') = \text{Tr} \left[ O_{\text{MMD}}^{(\sigma)}(\rho \otimes \rho') \right], \quad (\text{C4})$$

with the MMD observable defined as

$$O_{\text{MMD}}^{(\sigma)} := \sum_{\mathbf{x}, \mathbf{y}} K_{\sigma}(\mathbf{x}, \mathbf{y}) |\mathbf{x}\rangle\langle\mathbf{x}| \otimes |\mathbf{y}\rangle\langle\mathbf{y}|, \quad (\text{C5})$$

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

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

$$O_{\text{MMD}}^{(\sigma)} = \sum_{\mathbf{x}, \mathbf{y}} \bigotimes_{i=1}^n \left[ \left( \frac{\mathbb{1}_i + (-1)^{x_i} Z_i}{2} \right) \otimes \left( \frac{\mathbb{1}_{n+i} + (-1)^{y_i} Z_{n+i}}{2} \right) \exp \left( -\frac{(x_i - y_i)^2}{2\sigma} \right) \right] \quad (\text{C6})$$

$$= \bigotimes_{i=1}^n \sum_{x_i, y_i} \left[ \left( \frac{\mathbb{1}_i + (-1)^{x_i} Z_i}{2} \right) \otimes \left( \frac{\mathbb{1}_{n+i} + (-1)^{y_i} Z_{n+i}}{2} \right) \exp \left( -\frac{(x_i - y_i)^2}{2\sigma} \right) \right] \quad (\text{C7})$$

$$= \bigotimes_{i=1}^n [(1 - p_{\sigma})\mathbb{1}_i \otimes \mathbb{1}_{n+i} + p_{\sigma} Z_i \otimes Z_{n+i}] \quad (\text{C8})$$

$$= \sum_{A \subseteq \mathcal{N}} (1 - p_{\sigma})^{n-|A|} p_{\sigma}^{|A|} \bigotimes_{i \in A} (Z_i \otimes Z_{n+i}) \quad (\text{C9})$$

$$= \sum_{l=0}^n \binom{n}{l} (1 - p_{\sigma})^{n-l} p_{\sigma}^l D_{2l}, \quad (\text{C10})$$

where we denote

$$p_\sigma = (1 - e^{-1/(2\sigma)})/2, \quad (\text{C11})$$

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,

$$A \subseteq \mathcal{N} = \{1, 2, \dots, n\}. \quad (\text{C12})$$

For example for  $n = 3$ , we have

$$A \in \{\{\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}. \quad (\text{C13})$$

In the last step, we denote

$$D_{2l} = \frac{1}{\binom{n}{l}} \sum_{\substack{A \subseteq \mathcal{N} \\ |A|=l}} \bigotimes_{i \in A} (Z_i \otimes Z_{n+i}), \quad (\text{C14})$$

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

Interestingly,  $p_\sigma$  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,

$$w_\sigma(l) = \binom{n}{l} (1 - p_\sigma)^{n-l} p_\sigma^l, \quad (\text{C15})$$

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

$$(n+1)p_\sigma - 1 \leq l_{\max} = \arg \max(w_\sigma(l)) \leq (n+1)p_\sigma. \quad (\text{C16})$$

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

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] = 2np_\sigma, \quad (\text{C17})$$

and its variance is

$$\text{Var}_{l \sim w_\sigma(l)}[2l] = 4np_\sigma(1 - p_\sigma). \quad (\text{C18})$$

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

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

**Proposition 2** (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*

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] \in \Theta(n). \quad (\text{C19})$$

Similarly, the variance in the bodyness is given by

$$\text{Var}_{l \sim w_\sigma(l)}[2l] \in \Theta(n). \quad (\text{C20})$$

*Proof.* By considering that  $p_\sigma, (1 - p_\sigma) \in \mathcal{O}(1)$  for  $\sigma \in \mathcal{O}(1)$  in conjunction with Eqs. (C17) and (C18), 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  $k < l_{\max}$ )

$$\sum_{l=0}^k w_\sigma(l) = \sum_{l=0}^k \binom{n}{l} (1 - p_\sigma)^{n-l} p_\sigma^l \quad (\text{C21})$$

$$\leq (k+1) \binom{n}{k} (1 - p_\sigma)^{n-k} p_\sigma^k \quad (\text{C22})$$

$$\leq (k+1) \left(\frac{ne}{k}\right)^k (1 - p_\sigma)^{n-k} p_\sigma^k, \quad (\text{C23})$$

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  $k \in \mathcal{O}(1)$ ,

$$\sum_{l=0}^k w_\sigma(l) \in \mathcal{O}(\text{poly}(n)/b^n), \quad (\text{C24})$$

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

$$p(\sigma) = \frac{1 - \left(1 - \frac{1}{2\sigma} + \mathcal{O}\left(\frac{1}{\sigma^2}\right)\right)}{2} \quad (\text{C25})$$

$$= \frac{1}{4\sigma} + \mathcal{O}\left(\frac{1}{\sigma^2}\right) \quad (\text{C26})$$

$$\in \mathcal{O}\left(\frac{1}{n}\right) \quad (\text{C27})$$

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

$$\mathbb{E}_{l \sim w_\sigma(l)}[2l] \in \mathcal{O}(1) \quad (\text{C28})$$

$$\text{Var}_{l \sim w_\sigma(l)}[2l] \in \mathcal{O}(1). \quad (\text{C29})$$

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 3 in the main text, which is proven below.

**Proposition 3** (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  $O_{\text{MMD}}^{(\sigma)}$ ,*

$$\tilde{O}_{\text{MMD}}^{(\sigma,k)} := \sum_{l=0}^k w_\sigma(l) D_{2l}, \quad (\text{C30})$$

where  $w_\sigma(l)$  are Bernoulli-distributed weights defined in Eq. (31). For  $\sigma \in \Theta(n)$ , the difference between the exact and local approximation of the loss is bounded as

$$|\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\boldsymbol{\theta})| \leq \epsilon(k), \quad (\text{C31})$$

with

$$\epsilon(k) \in \mathcal{O}\left(n(c/k)^k\right), \quad (\text{C32})$$

for some positive constant  $c$ .

*Proof.* Let  $\rho_{\theta} = |\psi(\theta)\rangle\langle\psi(\theta)|$  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

$$\begin{aligned} |\mathcal{L}_{\text{MMD}}^{(\sigma)}(\theta) - \tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\theta)| &= \left| \text{Tr} \left[ \left( O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\theta} \otimes \rho_{\theta}) \right] - 2 \text{Tr} \left[ \left( O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\theta} \otimes \rho_{\tilde{p}}) \right] \right. \\ &\quad \left. + \text{Tr} \left[ \left( O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\tilde{p}} \otimes \rho_{\tilde{p}}) \right] \right| \end{aligned} \quad (\text{C33})$$

$$\begin{aligned} &\leq \left| \text{Tr} \left[ \left( O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\theta} \otimes \rho_{\theta}) \right] \right| + 2 \left| \text{Tr} \left[ \left( O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\theta} \otimes \rho_{\tilde{p}}) \right] \right| \\ &\quad + \left| \text{Tr} \left[ \left( O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right) (\rho_{\tilde{p}} \otimes \rho_{\tilde{p}}) \right] \right| \end{aligned} \quad (\text{C34})$$

$$\begin{aligned} &\leq \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \|\rho_{\theta} \otimes \rho_{\theta}\|_1 + 2 \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \|\rho_{\theta} \otimes \rho_{\tilde{p}}\|_1 \\ &\quad + \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \|\rho_{\tilde{p}} \otimes \rho_{\tilde{p}}\|_1 \end{aligned} \quad (\text{C35})$$

$$= 4 \left\| O_{\text{MMD}}^{(\sigma)} - \tilde{O}_{\text{MMD}}^{(\sigma,k)} \right\|_{\infty} \quad (\text{C36})$$

$$= 4 \left\| \sum_{l=k+1}^n w_{\sigma}(l) D_{2l} \right\|_{\infty} \quad (\text{C37})$$

$$\leq 4 \sum_{l=k+1}^n \binom{n}{l} (1-p_{\sigma})^{n-l} p_{\sigma}^l \quad (\text{C38})$$

$$\leq 4 \sum_{l=k+1}^n \binom{n}{l} \left( \frac{1-e^{-1/2\sigma}}{2} \right)^l \quad (\text{C39})$$

$$\leq 4 \sum_{l=k+1}^n \left( \frac{ne}{4l\sigma} \right)^l, \quad (\text{C40})$$

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  $f'(x) = f(x) \left[\ln\left(\frac{ne}{4\sigma x}\right) - 1\right]$  which leads to the maximum of  $f(x)$  at  $x^* = n/(4\sigma)$ . This leads to

$$\sum_{l=k+1}^n \left(\frac{ne}{4\sigma l}\right)^l \leq (n-k) \left(\frac{ne}{4\sigma k^*}\right)^{k^*}, \quad (\text{C41})$$

where  $k^* = \max(k, n/4\sigma)$ . Finally, if we assume that  $\sigma \in \Theta(1)$  and if  $k \geq n/(4\sigma)$ , then we obtain

$$\epsilon(k) \in \mathcal{O}\left(n\left(\frac{c}{k}\right)^k\right) \quad (\text{C42})$$

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

## 2. 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 and provide further discussion.

The variance of the MMD can be computed as

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}(\boldsymbol{\theta})] = \text{Var}_{\boldsymbol{\theta}} \left[ \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})K(\mathbf{x}, \mathbf{y}) - 2 \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}) + \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} p(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}) \right] \quad (\text{C43})$$

$$= \text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})] + 4\text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{p,q}(\boldsymbol{\theta})] - 4\text{Cov}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta}), \mathcal{K}_{p,q}(\boldsymbol{\theta})], \quad (\text{C44})$$

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

$$\mathcal{K}_{q,q}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})q_{\boldsymbol{\theta}}(\mathbf{y})K(\mathbf{x}, \mathbf{y}), \quad (\text{C45})$$

and

$$\mathcal{K}_{p,q}(\boldsymbol{\theta}) = \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\mathbf{x})p(\mathbf{y})K(\mathbf{x}, \mathbf{y}). \quad (\text{C46})$$

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

$$U(\boldsymbol{\theta}) = \bigotimes_{i=1}^n U_i(\boldsymbol{\theta}_i), \quad (\text{C47})$$

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

$$q_{\boldsymbol{\theta}}(\mathbf{x}) = \text{Tr} [U(\boldsymbol{\theta})|\mathbf{0}\rangle\langle\mathbf{0}|U^\dagger(\boldsymbol{\theta})|\mathbf{x}\rangle\langle\mathbf{x}|] \quad (\text{C48})$$

$$= \prod_{i=1}^n \text{Tr} [U_i(\boldsymbol{\theta}_i)|0_i\rangle\langle 0_i|U_i^\dagger(\boldsymbol{\theta}_i)|x_i\rangle\langle x_i|]. \quad (\text{C49})$$

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

### a. 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

$$\mathbb{E}_{\boldsymbol{\theta}}[q_{\boldsymbol{\theta}}(\mathbf{x})] = \int dU(\boldsymbol{\theta}) \text{Tr} [U(\boldsymbol{\theta})|\mathbf{0}\rangle\langle\mathbf{0}|U^\dagger(\boldsymbol{\theta})|\mathbf{x}\rangle\langle\mathbf{x}|] \quad (\text{C50})$$

$$= \prod_{i=1}^n \int dU_i(\boldsymbol{\theta}_i) \text{Tr} [U_i(\boldsymbol{\theta}_i)|0_i\rangle\langle 0_i|U_i^\dagger(\boldsymbol{\theta}_i)|x_i\rangle\langle x_i|] \quad (\text{C51})$$

$$= \prod_{i=1}^n \frac{\text{Tr}[|0_i\rangle\langle 0_i|] \text{Tr}[|x_i\rangle\langle x_i|]}{2} \quad (\text{C52})$$

$$= \frac{1}{2^n}, \quad (\text{C53})$$

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. [109])

$$\int dV V^{\otimes 1} |0\rangle \langle 0|^{\otimes 1} V^{\dagger \otimes 1} = \frac{1}{2} \mathbb{1} \quad (\text{C54})$$

$$\int dV V^{\otimes 2} |0\rangle \langle 0|^{\otimes 2} V^{\dagger \otimes 2} = \frac{1}{6} (\mathbb{1} \otimes \mathbb{1} + S_{12}) \quad (\text{C55})$$

$$\int dV V^{\otimes 3} |0\rangle \langle 0|^{\otimes 3} V^{\dagger \otimes 3} = \frac{1}{24} (\mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} + S_{12} + S_{13} + S_{23} + S_{23} S_{12} + S_{23} S_{13}), \quad (\text{C56})$$

$$\begin{aligned} \int dV V^{\otimes 4} |0\rangle \langle 0|^{\otimes 4} V^{\dagger \otimes 4} &= \frac{1}{120} (\mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} + S_{12} + S_{13} + S_{14} + S_{23} + S_{24} + S_{34} + S_{34} S_{12} + S_{24} S_{13} + S_{23} S_{14} \\ &\quad + S_{23} S_{12} + S_{24} S_{12} + S_{23} S_{13} + S_{34} S_{13} + S_{24} S_{14} + S_{34} S_{23} + S_{34} S_{24} + S_{34} S_{41} \\ &\quad + S_{34} S_{23} S_{12} + S_{34} S_{24} S_{12} + S_{24} S_{23} S_{13} + S_{24} S_{34} S_{13} + S_{23} S_{34} S_{14} + S_{23} S_{24} S_{14}) \end{aligned} \quad (\text{C57})$$

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*

$$O = \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma \sigma, \quad (\text{C58})$$

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

$$\text{Var}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}[O] = \sum_{\sigma \in \mathfrak{p}_n \setminus \{\mathbb{1}^{\otimes n}\}} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}}, \quad (\text{C59})$$

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

*Proof.* We consider the arbitrary observable  $O$  which can be decomposed into the Pauli basis as

$$O = \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma \sigma, \quad (\text{C60})$$

where the weights  $\lambda_\sigma$  are real constants and  $\mathfrak{p}_n = \{\mathbb{1}, X, Y, Z\}^{\otimes n}$  is the Pauli ensemble on  $n$  qubits. We denote  $s(\sigma)$  as a support of  $\sigma$  which is a subset of qubits that  $\sigma$  acts non-trivially on and  $|s(\sigma)|$  as a cardinality of  $s(\sigma)$ <sup>5</sup>.

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

$$\text{Var}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}[O] = \mathbb{E}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} [|\langle \psi | O | \psi \rangle|^2] - \left( \mathbb{E}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} [|\langle \psi | O | \psi \rangle|] \right)^2 \quad (\text{C61})$$

$$= \langle O^2 \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} - \langle O \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}^2. \quad (\text{C62})$$

---

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

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

$$\langle O \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} = \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma \langle \sigma \rangle_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}} \quad (\text{C63})$$

$$= \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma \prod_{i=1}^n \langle \sigma_i \rangle_{|\psi_i\rangle \sim \text{Haar}_1} \quad (\text{C64})$$

$$= \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma \prod_{i=1}^n \delta(\sigma_i = \mathbb{1}) \quad (\text{C65})$$

$$= \lambda_{\mathbb{1}^{\otimes n}}, \quad (\text{C66})$$

where, in the third equality, we use the Haar integration formula in Eq. (C54) 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. [72] to integrate over the random product states but replace the unitary  $\tilde{U}^\dagger W \hat{U}$  in Ref. [72] with an observable  $O$ . This directly leads to

$$\langle O^2 \rangle = \frac{1}{6^n} \sum_{A \subseteq \mathcal{N}} \text{Tr}[O_A^2], \quad (\text{C67})$$

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. (C12).

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

$$\sigma_A = \text{Tr}_{\bar{A}} \left[ \bigotimes_{i=1}^n \sigma_i \right] \quad (\text{C68})$$

$$= \left[ \prod_{i \notin A} \text{Tr}[\sigma_i] \right] \cdot \left[ \bigotimes_{i \in A} \sigma_i \right] \quad (\text{C69})$$

$$= 2^{n-|A|} \delta(s(\sigma) \subseteq A) \bigotimes_{i \in A} \sigma_i, \quad (\text{C70})$$

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  $\sigma$  acts non-trivially is a subset of  $A$ .

Now, we consider

$$\text{Tr}[O_A^2] = \text{Tr}_A [\text{Tr}_{\bar{A}}[O] \cdot \text{Tr}_{\bar{A}}[O]] \quad (\text{C71})$$

$$= \text{Tr} \left[ \left( \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma 2^{n-|A|} \delta(s(\sigma) \subseteq A) \bigotimes_{i \in A} \sigma_i \right) \left( \sum_{\sigma' \in \mathfrak{p}_n} \lambda_{\sigma'} 2^{n-|A|} \delta(s(\sigma') \subseteq A) \bigotimes_{i \in A} \sigma'_i \right) \right] \quad (\text{C72})$$

$$= \sum_{\sigma, \sigma' \in \mathfrak{p}_n} \lambda_\sigma \lambda_{\sigma'} 2^{2(n-|A|)} \delta(s(\sigma) \subseteq A) \delta(s(\sigma') \subseteq A) \left( \prod_{i \in A} \text{Tr}[\sigma_i \sigma'_i] \right) \quad (\text{C73})$$

$$= \sum_{\sigma, \sigma' \in \mathfrak{p}_n} \lambda_\sigma \lambda_{\sigma'} 2^{2(n-|A|)} \delta(s(\sigma) \subseteq A) \delta(s(\sigma') \subseteq A) \left( 2^{|A|} \prod_{i \in A} \delta(\sigma_i = \sigma'_i) \right) \quad (\text{C74})$$

$$= \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 2^{2n-|A|} \delta(s(\sigma) \subseteq A), \quad (\text{C75})$$

where the third equality is due to  $\text{Tr}[(\otimes_{i \in A} \sigma_i)(\otimes_{i \in A} \sigma'_i)] = \prod_{i \in A} \text{Tr}[\sigma_i \sigma'_i]$ , 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  $\sigma$  acts non-trivially only on  $A$  (i.e.,  $\delta(s(\sigma) \in A)$ ) together with the reduced Pauli strings on  $A$  are the same for  $\sigma$  and  $\sigma'$  (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

$$\langle O^2 \rangle = \frac{1}{6^n} \sum_{A \subseteq \mathcal{N}} \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 2^{2n-|A|} \delta(s(\sigma) \subseteq A) \quad (\text{C76})$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 \sum_{A \subseteq \mathcal{N}} 2^{-|A|} \delta(s(\sigma) \subseteq A) \quad (\text{C77})$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 \sum_{A' \subseteq \mathcal{N} \setminus s(\sigma)} 2^{-|s(\sigma)| - |A'|} \quad (\text{C78})$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 2^{-|s(\sigma)|} \sum_{|A'|=0}^{n-|s(\sigma)|} \binom{n-|s(\sigma)|}{|A'|} 2^{-|A'|} \quad (\text{C79})$$

$$= \left(\frac{2}{3}\right)^n \sum_{\sigma \in \mathfrak{p}_n} \lambda_\sigma^2 2^{-|s(\sigma)|} \left(\frac{3}{2}\right)^{n-|s(\sigma)|} \quad (\text{C80})$$

$$= \sum_{\sigma \in \mathfrak{p}_n} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}}, \quad (\text{C81})$$

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

$$\text{Var}_{|\psi\rangle \sim \text{Haar}_1^{\otimes n}}[O] = \sum_{\sigma \in \mathfrak{p}_n} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}} - (\lambda_{\mathbb{1}^{\otimes n}})^2 \quad (\text{C82})$$

$$= \sum_{\sigma \in \mathfrak{p}_n \setminus \mathbb{1}^{\otimes n}} \frac{\lambda_\sigma^2}{3^{|s(\sigma)|}}, \quad (\text{C83})$$

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

### b. 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), which uses the classical Gaussian kernel as defined in Eq. (12) with the bandwidth  $\sigma$ , and a quantum generative model that is comprised of a tensor-product ansatz  $U = \otimes_i^n U_i(\theta_i)$  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

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{L}_{\text{MMD}}^{(\sigma)}(\boldsymbol{\theta})] = B_\sigma + 4C_\sigma(\tilde{P}), \quad (\text{C84})$$

with

$$B_\sigma = \left[ \frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right]^n - \left[ \frac{4 + 4e^{-1/2\sigma} + e^{-1/\sigma}}{9} \right]^n, \quad (\text{C85})$$

and

$$C_\sigma(\tilde{P}) = \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (1 - p_\sigma)^{2(n-|A|)} \left( \frac{p_\sigma^2}{3} \right)^{|A|} z_A^2(\tilde{P}), \quad (\text{C86})$$

where  $p_\sigma = (1 - e^{-\frac{1}{2\sigma}})/2$ ,  $\mathcal{N} = \{1, 2, \dots, n\}$ ,  $z_A(\tilde{P}) = \text{Tr}[(\otimes_{i \in A} Z_i) \rho_{\tilde{P}}] = \sum_{\mathbf{y}} \tilde{p}(\mathbf{y}) (-1)^{\sum_{i \in A} y_i}$  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. (C86) is over all possible subsets of  $\mathcal{N}$  excluding the empty set  $\{\}$ .

We remark that  $B_\sigma$  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. 5.

*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  $\mathcal{K}_{q,q}(\boldsymbol{\theta})$ :

$$\text{Var}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}^2(\boldsymbol{\theta})] - (\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{K}_{q,q}(\boldsymbol{\theta})])^2 \quad (\text{C87})$$

$$= \mathbb{E}_{\boldsymbol{\theta}} \left[ \left( \sum_{\mathbf{x}, \mathbf{y}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) \right)^2 \right] - \left( \mathbb{E}_{\boldsymbol{\theta}} \left[ \sum_{\mathbf{x}, \mathbf{y}} q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) \right] \right)^2 \quad (\text{C88})$$

$$= \sum_{\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'} \mathbb{E}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y}) q_{\boldsymbol{\theta}}(\mathbf{x}') q_{\boldsymbol{\theta}}(\mathbf{y}')] K(\mathbf{x}, \mathbf{y}) K(\mathbf{x}', \mathbf{y}') - \left( \sum_{\mathbf{x}, \mathbf{y}} \mathbb{E}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y})] K(\mathbf{x}, \mathbf{y}) \right)^2. \quad (\text{C89})$$

We now can express each individual model probability as in Eq. (C49) and then average over the parameters  $\boldsymbol{\theta}$ . This requires us to perform Haar integration for the first and second terms in Eq. (C89), respectively.

First, consider

$$\sum_{\mathbf{x}, \mathbf{y}} \mathbb{E}_{\boldsymbol{\theta}} [q_{\boldsymbol{\theta}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\mathbf{y})] K(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \int dU_i(\boldsymbol{\theta}_i) \text{Tr} \left[ (U_i(\boldsymbol{\theta}_i))^{\otimes 2} |0_i\rangle\langle 0_i|^{\otimes 2} (U_i^\dagger(\boldsymbol{\theta}_i))^{\otimes 2} (|x_i\rangle\langle x_i| \otimes |y_i\rangle\langle y_i|) \right] K(\mathbf{x}, \mathbf{y}) \quad (\text{C90})$$

$$= \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \text{Tr} \left[ \left( \int dU_i(\boldsymbol{\theta}_i) (U_i(\boldsymbol{\theta}_i))^{\otimes 2} |0_i\rangle\langle 0_i|^{\otimes 2} (U_i^\dagger(\boldsymbol{\theta}_i))^{\otimes 2} \right) (|x_i\rangle\langle x_i| \otimes |y_i\rangle\langle y_i|) \right] K(\mathbf{x}, \mathbf{y}) \quad (\text{C91})$$

$$= \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \text{Tr} \left[ \left( \frac{\mathbb{1} \otimes \mathbb{1} + S_{12}}{6} \right) (|x_i\rangle\langle x_i| \otimes |y_i\rangle\langle y_i|) \right] e^{-\frac{(x_i - y_i)^2}{2\sigma}} \quad (\text{C92})$$

$$= \sum_{\mathbf{x}, \mathbf{y}} \prod_{i=1}^n \left( \frac{1 + \delta_{x_i, y_i}}{6} \right) e^{-\frac{(x_i - y_i)^2}{2\sigma}} \quad (\text{C93})$$

$$= \sum_{\mathbf{x}} \prod_{i=1}^n \left[ \left( \frac{1 + \delta_{x_i, 0}}{6} \right) e^{-\frac{(x_i)^2}{2\sigma}} + \left( \frac{1 + \delta_{x_i, 1}}{6} \right) e^{-\frac{(x_i - 1)^2}{2\sigma}} \right] \quad (\text{C94})$$

$$= \left( \frac{2 + e^{-1/2\sigma}}{3} \right)^n, \quad (\text{C95})$$

where, in the third equality, we use Eq. (C55), 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. (C89) can be computed via Haar integration using Eq. (C57) 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

$$\sum_{\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'} \mathbb{E}_{\theta} [q_{\theta}(\mathbf{x})q_{\theta}(\mathbf{y})q_{\theta}(\mathbf{x}')q_{\theta}(\mathbf{y}')] K(\mathbf{x}, \mathbf{y})K(\mathbf{x}', \mathbf{y}') = \left( \frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right)^n. \quad (\text{C96})$$

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

$$\text{Var}_{\theta} [\mathcal{K}_{p,q}(\theta)] = \left( \frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right)^n - \left( \frac{4 + 4e^{-1/2\sigma} + e^{-1/\sigma}}{9} \right)^n. \quad (\text{C97})$$

(ii) Computing the variance of  $\mathcal{K}_{p,q}(\theta)$ : There are two alternative ways of doing this, leading to two equivalent expressions of the variance of  $\mathcal{K}_{p,q}(\theta)$ . 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,  $\mathcal{K}_{p,q}(\theta) = \text{Tr} [U(\theta)|0\rangle\langle 0|U^{\dagger}(\theta)O_{\tilde{p}}^{(\sigma)}(\tilde{P})]$ . 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

$$O_{\tilde{p}}^{(\sigma)} = \sum_{A \subseteq \mathcal{N}} (1 - p_{\sigma})^{n-|A|} p_{\sigma}^{|A|} z_A(\tilde{P}) \bigotimes_{i \in A} Z_i, \quad (\text{C98})$$

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

$$z_A(\tilde{P}) = \text{Tr} \left[ \left( \bigotimes_{i \in A} Z_i \right) \rho_{\tilde{p}} \right] \quad (\text{C99})$$

$$= \sum_{\mathbf{y}} \tilde{p}(\mathbf{y}) (-1)^{\sum_{i \in A} y_i}, \quad (\text{C100})$$

where  $\rho_{\tilde{p}}$  is the quantum state associated with the training data with  $\tilde{p}(\mathbf{x}) = \text{Tr}[\rho_{\tilde{p}}|\mathbf{x}\rangle\langle\mathbf{x}|]$ <sup>6</sup>.

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

$$\text{Var}_{\theta} [\mathcal{K}_{p,q}(\theta)] = \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (1 - p_{\sigma})^{2(n-|A|)} \left( \frac{p_{\sigma}^2}{3} \right)^{|A|} z_A^2(\tilde{P}). \quad (\text{C101})$$

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  $\mathcal{K}_{p,p}(\theta)$  and  $\mathcal{K}_{p,q}(\theta)$ : By direct computation as in (i), we have

$$\text{Cov}_{\theta} [\mathcal{K}_{q,q}(\theta), \mathcal{K}_{p,q}(\theta)] = \mathbb{E}_{\theta} [\mathcal{K}_{q,q}(\theta)\mathcal{K}_{p,q}(\theta)] - \mathbb{E}_{\theta} [\mathcal{K}_{q,q}(\theta)]\mathbb{E}_{\theta} [\mathcal{K}_{p,q}(\theta)] \quad (\text{C102})$$

$$= \sum_{\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}'} p(\mathbf{y}') \left( \mathbb{E}_{\theta} [q_{\theta}(\mathbf{x})q_{\theta}(\mathbf{y})q_{\theta}(\mathbf{y}')] - \mathbb{E}_{\theta} [q_{\theta}(\mathbf{x})q_{\theta}(\mathbf{y})]\mathbb{E}_{\theta} [q_{\theta}(\mathbf{y}')] \right) K(\mathbf{x}, \mathbf{y})K(\mathbf{x}', \mathbf{y}') \quad (\text{C103})$$

$$= 0, \quad (\text{C104})$$

<sup>6</sup> 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. (C10). That is,  $O_{\tilde{p}}^{(\sigma)} = \text{Tr}_1 \left[ (\mathbb{1} \otimes \rho_{\tilde{p}}) O_{\text{MMD}}^{(\sigma)} \right]$ .

where the last equality follows from

$$\sum_{\mathbf{x}, \mathbf{y}} \left( \mathbb{E}_{\theta} [q_{\theta}(\mathbf{x})q_{\theta}(\mathbf{y})q_{\theta}(\mathbf{y}')] - \mathbb{E}_{\theta} [q_{\theta}(\mathbf{x})q_{\theta}(\mathbf{y})] \mathbb{E}_{\theta} [q_{\theta}(\mathbf{x}')] \right) K(\mathbf{x}, \mathbf{y}) = 0 \quad (\text{C105})$$

which holds for any  $\mathbf{x}'$  and  $\mathbf{y}'$  from Eq. (C50), Eq. (C55) and Eq. (C56).

By substituting Eq. (C97), Eq. (C101) and Eq. (C104) back into the MMD variance expression in Eq. (C44), the proof is completed.  $\square$

*c. Variance scaling and trainability of MMD*

We now analyze how the scaling of the variance depends on the bandwidth  $\sigma$ . 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 is stated below.

**Theorem 2** (Product ansatz trainability of MMD, formal). *Consider the MMD loss function  $\mathcal{L}_{\text{MMD}}^{(\sigma)}(\theta)$  as defined in Eq. (11), which uses the classical Gaussian kernel as defined in Eq. (12) 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  $\sigma$ .*

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

$$\text{Var}_{\theta} [\mathcal{L}_{\text{MMD}}^{(\sigma)}(\theta)] \in \mathcal{O}(1/b^n), \quad (\text{C106})$$

with some  $b > 1$ .

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

$$\text{Var}_{\theta} [\mathcal{L}_{\text{MMD}}^{(\sigma)}(\theta)] = B_{\sigma} + 4C_{\sigma}(\tilde{P}), \quad (\text{C107})$$

where the variance of the first term  $B_{\sigma}$  is lower-bounded as

$$B_{\sigma} \in \Omega(1/n), \quad (\text{C108})$$

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

$$C_{\sigma}(\tilde{P}) \in \Omega(1/\text{poly}(n)), \quad (\text{C109})$$

provided that

$$\sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}) \in \Omega(1/\text{poly}(n)), \quad (\text{C110})$$

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. (C84) for two scenarios of the bandwidth values.

(i) For  $\sigma \in \mathcal{O}(1)$ : We show that both  $B_\sigma$  and  $C_\sigma(\tilde{P})$  in the MMD variance are exponentially small. First, from Eq. (C85) we know that

$$B_\sigma \leq \left[ \frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right]^n. \quad (\text{C111})$$

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. (C86), we upper bound as

$$C_\sigma(\tilde{P}) \leq \sum_{\substack{A \subseteq \mathcal{N} \\ A \neq \{\}}} (1 - p_\sigma)^{2(n-|A|)} \left( \frac{p_\sigma^2}{3} \right)^{|A|} \quad (\text{C112})$$

$$= \sum_{|A|=1}^n \binom{n}{|A|} (1 - p_\sigma)^{2(n-|A|)} \left( \frac{p_\sigma^2}{3} \right)^{|A|} \quad (\text{C113})$$

$$\leq \sum_{|A|=0}^n \binom{n}{|A|} (1 - p_\sigma)^{2(n-|A|)} \left( \frac{p_\sigma^2}{3} \right)^{|A|} \quad (\text{C114})$$

$$= \left( (1 - p_\sigma)^2 + \frac{p_\sigma^2}{3} \right)^n \quad (\text{C115})$$

$$= \left[ \frac{1 + e^{-1/2\sigma} + e^{-1/\sigma}}{3} \right]^n, \quad (\text{C116})$$

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  $p_\sigma$  from Eq. (C11). 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

$$\text{Var}_\theta[\mathcal{L}_{\text{MMD}}(\theta)] \in \mathcal{O}(1/b^n), \quad (\text{C117})$$

for some  $b > 0$ .

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

$$B_\sigma = \left[ \frac{7 + 6e^{-1/2\sigma} + 2e^{-1/\sigma}}{15} \right]^n - \left[ \frac{4 + 4e^{-1/2\sigma} + e^{-1/\sigma}}{9} \right]^n \quad (\text{C118})$$

$$\geq \left[ \frac{7 + 6 \left( 1 - \frac{1}{2\sigma} + \frac{1}{8\sigma^2} - \frac{1}{48\sigma^3} \right) + 2 \left( 1 - \frac{1}{\sigma} + \frac{1}{2\sigma^2} - \frac{1}{6\sigma^3} \right)}{15} \right]^n - \left[ \frac{4 + 4 \left( 1 - \frac{1}{2\sigma} + \frac{1}{8\sigma^2} \right) + \left( 1 - \frac{1}{\sigma} + \frac{1}{2\sigma^2} \right)}{9} \right]^n \quad (\text{C119})$$

$$= \left[ 1 - \frac{1}{3\sigma} + \frac{7}{60\sigma^2} - \frac{11}{360\sigma^3} \right]^n - \left[ 1 - \frac{1}{3\sigma} + \frac{1}{9\sigma^2} \right]^n \quad (\text{C120})$$

$$= \left( 1 - \frac{1}{3\sigma} \right)^n \left[ \left[ 1 + \frac{\frac{7}{60\sigma^2} - \frac{11}{360\sigma^3}}{1 - \frac{1}{3\sigma}} \right]^n - \left[ 1 + \frac{\frac{1}{9\sigma^2}}{1 - \frac{1}{3\sigma}} \right]^n \right] \quad (\text{C121})$$

$$\geq \left( 1 - \frac{1}{3\sigma} \right)^{n-1} \left( \frac{n}{180\sigma^2} \right) \left( 1 - \frac{11}{2\sigma} \right) \quad (\text{C122})$$

$$= \left( \frac{n}{180\sigma^2} \right) \left( 1 - \frac{11}{2\sigma} \right) \left[ \left( 1 - \frac{1}{3\sigma} \right)^{3\sigma} \right]^{(n-1)/3\sigma}, \quad (\text{C123})$$

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*

$$f(x) \geq \frac{1}{e} \left( 1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)x^j} \right). \quad (\text{C124})$$

*Proof.* We consider

$$f(x) = \exp(x \ln(1 - 1/x)) \quad (\text{C125})$$

$$= \exp \left( x \left( \sum_{j=1}^{\infty} -\frac{1}{jx^j} \right) \right) \quad (\text{C126})$$

$$= \frac{1}{e} \cdot \exp \left( \sum_{j=1}^{\infty} -\frac{1}{(j+1)x^j} \right) \quad (\text{C127})$$

$$\geq \frac{1}{e} \left( 1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)x^j} \right), \quad (\text{C128})$$

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

$$B_\sigma \geq \left( \frac{n}{180\sigma^2} \right) \left( 1 - \frac{11}{2\sigma} \right) \left[ \frac{1}{e} \left( 1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)(3\sigma)^j} \right) \right]^{(n-1)/3\sigma}, \quad (\text{C129})$$

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

Similarly, for the second term, we have

$$C_\sigma(\tilde{P}) = \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}}} (1 - p_\sigma)^{2(n-|A|)} \left(\frac{p_\sigma^2}{3}\right)^{|A|} z_A^2(\tilde{P}) \quad (\text{C130})$$

$$= \left[ \frac{1 + 2e^{-1/2\sigma} + e^{-1/\sigma}}{4} \right]^n \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}}} \left(\frac{\tanh^2(1/4\sigma)}{3}\right)^{|A|} z_A^2(\tilde{P}) \quad (\text{C131})$$

$$\geq \left[ \frac{1 + 2\left(1 - \frac{1}{2\sigma}\right) + \left(1 - \frac{1}{\sigma}\right)}{4} \right]^n \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}}} \left(\frac{\frac{1}{16\sigma^2} \left(1 - \frac{1}{24\sigma^2}\right)}{3}\right)^{|A|} z_A^2(\tilde{P}) \quad (\text{C132})$$

$$\geq \left[ \left(1 - \frac{1}{2\sigma}\right)^{2\sigma} \right]^{\frac{n}{2\sigma}} \left(\frac{1}{48\sigma^2} \left(1 - \frac{1}{24\sigma^2}\right)\right)^k \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}) \quad (\text{C133})$$

$$\geq \left[ \frac{1}{e} \left(1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)(2\sigma)^j}\right) \right]^{\frac{n}{2\sigma}} \left(\frac{1}{48\sigma^2} \left(1 - \frac{1}{24\sigma^2}\right)\right)^k \sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}), \quad (\text{C134})$$

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

$$\sum_{\substack{A \subseteq \mathcal{N} \\ ; A \neq \{\}, |A| \leq k}} z_A^2(\tilde{P}) \in \Omega(1/\text{poly}(n)), \quad (\text{C135})$$

the MMD variance is lower bounded as

$$\text{Var}_\theta[\mathcal{L}_{\text{MMD}}(\theta)] \in \Omega(1/n), \quad (\text{C136})$$

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

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. (C99)

$$z_A(\tilde{P}) = \text{Tr} \left[ \left( \bigotimes_{i \in A} Z_i \right) \rho_{\tilde{p}} \right] \quad (\text{C137})$$

$$= \sum_{\mathbf{y}} \tilde{p}(\mathbf{y}) (-1)^{\sum_{i \in A} y_i}, \quad (\text{C138})$$

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. (C12)), 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.

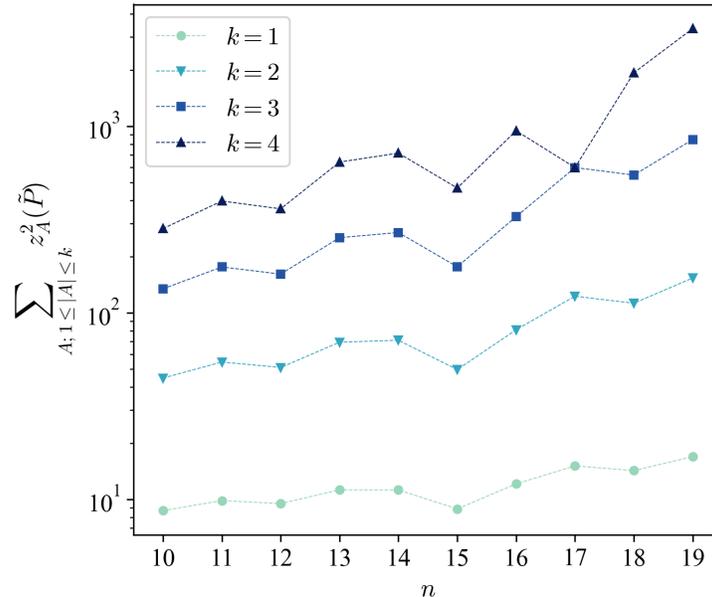


Figure 10. We plot  $\sum_{A \subseteq \mathcal{N}} z_A^2(\tilde{P})$  such that  $|A| \leq k$  and  $A \neq \{\}$  as a function of  $n$  and shows different  $k$  values. The datasets considered here corresponds to the real high energy physics dataset used in Sec. IV. Values of  $n = 10, 11, \dots, 19$  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.

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 Section IV. Fig. 10 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.

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

Our results so far (Theorem 2 and Conjecture 1) 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 Appendix C1, 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 3 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 Appendix C4 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

$$\tilde{p}(\mathbf{x}_A) = \sum_{\mathbf{x}_{\bar{A}} \in \{0,1\}^{\otimes(n-|A|)}} \tilde{p}(\mathbf{x}_A, \mathbf{x}_{\bar{A}}) \quad (\text{C139})$$

$$= \text{Tr} [\rho_{\tilde{p}} (|\mathbf{x}_A\rangle\langle\mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] , \quad (\text{C140})$$

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

$$q_{\theta}(\mathbf{x}_A) = \sum_{\mathbf{x}_{\bar{A}} \in \{0,1\}^{\otimes(n-|A|)}} q_{\theta}(\mathbf{x}_A, \mathbf{x}_{\bar{A}}) \quad (\text{C141})$$

$$= \text{Tr} [\rho_{\theta} (|\mathbf{x}_A\rangle\langle\mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] . \quad (\text{C142})$$

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

$$\text{Diag}(\text{Tr}_{\bar{A}}[\rho_{\tilde{p}}]) = \text{Diag}(\text{Tr}_{\bar{A}}[\rho_{\theta}]) , \quad (\text{C143})$$

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

In addition, we recall that the truncated version of the MMD observable defined in Eq. (C10) is of the form

$$\tilde{O}_{\text{MMD}}^{(\sigma,k)} = \sum_{l=0}^k \binom{n}{l} (1-p_{\sigma})^{n-l} p_{\sigma}^l D_{2l} \quad (\text{C144})$$

$$= \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} (1-p_{\sigma})^{n-|A|} p_{\sigma}^{|A|} \bigotimes_{i \in A} (Z_i \otimes Z_{n+i}) , \quad (\text{C145})$$

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

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\theta) = \text{Tr} \left[ \tilde{O}_{\text{MMD}}^{(\sigma,k)} (\rho_{\theta} \otimes \rho_{\theta}) \right] - 2 \text{Tr} \left[ \tilde{O}_{\text{MMD}}^{(\sigma,k)} (\rho_{\theta} \otimes \rho_{\tilde{p}}) \right] + \text{Tr} \left[ \tilde{O}_{\text{MMD}}^{(\sigma,k)} (\rho_{\tilde{p}} \otimes \rho_{\tilde{p}}) \right] \quad (\text{C146})$$

$$= \text{Tr} \left[ \tilde{O}_{\text{MMD}}^{(\sigma,k)} (\rho_{\theta} - \rho_{\tilde{p}})^{\otimes 2} \right] . \quad (\text{C147})$$

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

**Proposition 4** (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*

$$q_{\theta}(\mathbf{x}_A) = \tilde{p}(\mathbf{x}_A) , \quad (\text{C148})$$

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

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\theta) = 0 . \quad (\text{C149})$$

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

$$q_{\theta}(\mathbf{x}_B) \neq \tilde{p}(\mathbf{x}_B) . \quad (\text{C150})$$

*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. (C147) leading to

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\boldsymbol{\theta}) = \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} (1 - p_\sigma)^{n-|A|} p_\sigma^{|A|} [\langle Z_A \rangle_{\boldsymbol{\theta}} - \langle Z_A \rangle_{\tilde{p}}]^2, \quad (\text{C151})$$

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

$$Z_A = \sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} |\mathbf{x}_A\rangle \langle \mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}}. \quad (\text{C152})$$

So, the expectation of  $Z_A$  can be written as a sum of the marginals probabilities on  $A$  with the definition in Eq. (C139) as follows,

$$\langle Z_A \rangle_{\tilde{p}} = \sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} \tilde{p}(\mathbf{x}_A), \quad (\text{C153})$$

and

$$\langle Z_A \rangle_{\boldsymbol{\theta}} = \sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} q_{\boldsymbol{\theta}}(\mathbf{x}_A). \quad (\text{C154})$$

Then, we have

$$\tilde{\mathcal{L}}_{\text{MMD}}^{(\sigma,k)}(\boldsymbol{\theta}) = \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} (1 - p_\sigma)^{n-|A|} p_\sigma^{|A|} \left[ \sum_{\mathbf{x}_A} (-1)^{\sum_{i \in A} x_i} (\tilde{p}(\mathbf{x}_A) - q_{\boldsymbol{\theta}}(\mathbf{x}_A)) \right]^2 \quad (\text{C155})$$

$$= 0, \quad (\text{C156})$$

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 4, 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 (C151) and notice that the loss is minimized and equals to 0 if and only if

$$\langle Z_A \rangle_{\boldsymbol{\theta}} = \langle Z_A \rangle_{\tilde{p}}, \quad (\text{C157})$$

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

$$|\mathbf{x}_A\rangle \langle \mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}} = \bigotimes_{i \in A} \frac{1}{2} (\mathbb{1}_i + (-1)^{x_i} Z_i) \quad (\text{C158})$$

$$= \frac{1}{2^{|A|}} \sum_{B \subseteq A} (-1)^{\sum_{i \in B} x_i} Z_B, \quad (\text{C159})$$

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

$$\tilde{p}(\mathbf{x}_A) = \frac{1}{2^{|A|}} \sum_{B \subseteq A} (-1)^{\sum_{i \in B} x_i} \langle Z_B \rangle_{\tilde{p}}. \quad (\text{C160})$$

It is clear from Eq. (C151) that training on the  $k$ -truncated MMD learns all average parities of the target distribution up to  $k$  bits, Eq. (C157), and hence Eq. (C160) 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

$$\tilde{p}(\mathbf{x}_A) - q_{\theta}(\mathbf{x}_A) = \frac{1}{2^{|A|}} \sum_{B \subseteq A} (-1)^{\sum_{i \in B} x_i} (\langle Z_B \rangle_{\tilde{p}} - \langle Z_B \rangle_{\theta}), \quad (\text{C161})$$

$$= 0, \quad (\text{C162})$$

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

#### 4. Distinguishing marginals using an arbitrary loss

In Appendix C3, 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,

$$O = \sum_{\mathbf{x} \in \mathcal{X}} D_{\alpha}(\mathbf{x}) |\mathbf{x}\rangle \langle \mathbf{x}|. \quad (\text{C163})$$

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  $\alpha$ . 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

$$O^{(k)} = \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} c_A Z_A, \quad (\text{C164})$$

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  $\rho_{\theta_1}$ , 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  $\rho_{\theta_2}$ , matches the training distribution up to  $k$ -bit marginals but disagrees beyond. Both are characterized by the property

$$\text{Tr}[\rho_{\theta_1} (|\mathbf{x}_A\rangle \langle \mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] = \text{Tr}[\rho_{\theta_2} (|\mathbf{x}_A\rangle \langle \mathbf{x}_A| \otimes \mathbb{1}_{\bar{A}})] = \tilde{p}(\mathbf{x}_A). \quad (\text{C165})$$

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,

$$\langle Z_A \rangle_{\theta_1} = \langle Z_A \rangle_{\theta_2} = \langle Z_A \rangle_{\tilde{p}}. \quad (\text{C166})$$

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

$$d(\rho, \rho') = \mathcal{L}(\rho) - \mathcal{L}(\rho') \quad (\text{C167})$$

$$= \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} c_A (\text{Tr}_A[Z_A \rho] - \text{Tr}_A[Z_A \rho']) \quad (\text{C168})$$

$$= 0. \quad (\text{C169})$$

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  $D_{\alpha}$  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

$$O = \sum_{\substack{A_1, \dots, A_m \subseteq \mathcal{N} \\ |A_1|, \dots, |A_m| \leq k}} c_{A_1, \dots, A_m}(\alpha, \tilde{P}) (Z_{A_1} \otimes Z_{A_2} \otimes \dots \otimes Z_{A_m}), \quad (\text{C170})$$

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

This form of the general loss covers loss functions for quantum circuit Born machines (in particular the MMD, as outlined in Sec. III B), quantum GANS, quantum Boltzmann machines [110], 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.

#### Appendix D: Analysis on the quantum fidelity loss

In this section, we present an approach that can be used to estimate the local fidelity quantity  $\mathcal{L}_{QF}^{(L)}(\theta)$  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

$$\mathcal{L}_{QF}(\theta) = 1 - |\langle\phi|\psi(\theta)\rangle|^2. \quad (\text{D1})$$

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 [92]. Specifically, the local quantum fidelity is of the form

$$\mathcal{L}_{QF}^{(L)}(\theta) = 1 - \langle\phi|U(\theta)H_LU^\dagger(\theta)|\phi\rangle, \quad (\text{D2})$$

with

$$H_L = \frac{1}{n} \sum_{i=1}^n |0_i\rangle\langle 0_i| \otimes \mathbb{1}_i, \quad (\text{D3})$$

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

$$H_L = \frac{1}{n} \sum_{i=1}^n \left( \frac{\mathbb{1}_i + Z_i}{2} \right) \otimes \mathbb{1}_{\bar{i}} \quad (\text{D4})$$

$$= \frac{\mathbb{1}}{2} + \frac{1}{2n} \sum_i Z_i . \quad (\text{D5})$$

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

$$\mathcal{L}_{QF}^{(L)}(\boldsymbol{\theta}) = 1 - \langle \phi | U(\boldsymbol{\theta}) \left( \frac{\mathbb{1}}{2} + \frac{1}{2n} \sum_i Z_i \right) U^\dagger(\boldsymbol{\theta}) | \phi \rangle \quad (\text{D6})$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_i \langle \phi | U(\boldsymbol{\theta}) Z_i U^\dagger(\boldsymbol{\theta}) | \phi \rangle \quad (\text{D7})$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_{i, \mathbf{x}, \mathbf{x}'} \sqrt{\tilde{p}(\mathbf{x})\tilde{p}(\mathbf{x}')} \langle \mathbf{x} | U(\boldsymbol{\theta}) Z_i U^\dagger(\boldsymbol{\theta}) | \mathbf{x}' \rangle \quad (\text{D8})$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_{i, \mathbf{x}, \mathbf{x}'} \sqrt{\tilde{p}(\mathbf{x})\tilde{p}(\mathbf{x}')} \langle \mathbf{x} | U(\boldsymbol{\theta}) Z_i U^\dagger(\boldsymbol{\theta}) U_{\mathbf{x}', \mathbf{x}} | \mathbf{x} \rangle \quad (\text{D9})$$

$$= \frac{1}{2} - \frac{1}{2n} \sum_{i, \mathbf{x}, \mathbf{x}'} \sqrt{\tilde{p}(\mathbf{x})\tilde{p}(\mathbf{x}')} \langle \mathbf{x} | \tilde{U}(\boldsymbol{\theta}, i, \mathbf{x}, \mathbf{x}') | \mathbf{x} \rangle , \quad (\text{D10})$$

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 [111]. 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 [93] which allows us to stochastically optimize the loss in an unbiased manner.

### Appendix E: 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. IV, 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. 11 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  $\infty$ ) 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 not been discussed for quantum generative models before.

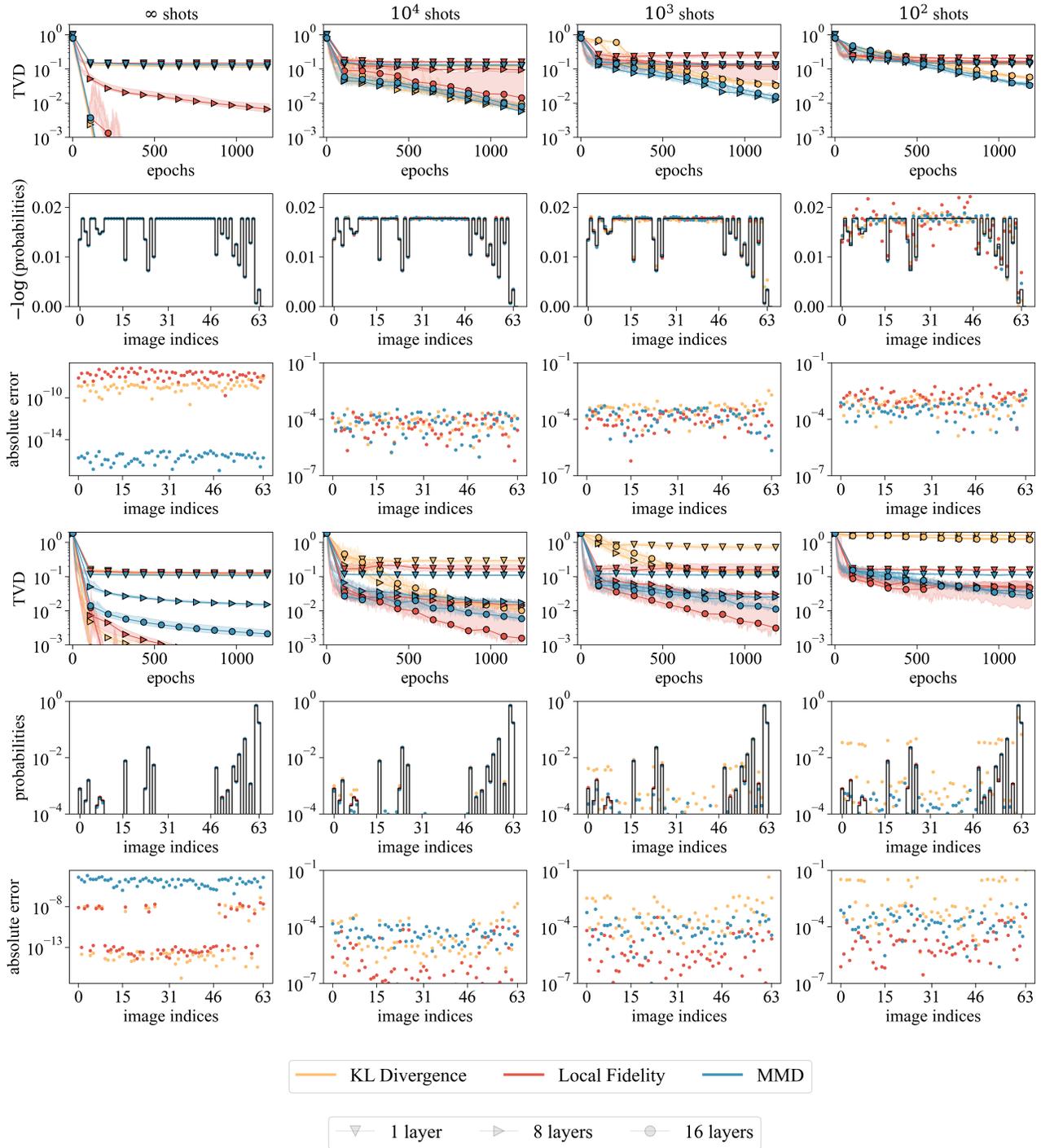


Figure 11. Training on the ECAL dataset with  $n = 6$  qubits on a dataset with exponential support (top three rows) and one with polynomial support (last three rows). The first rows shows the TVD during training, while the second displays the generated distribution against the target one (black) and the third the absolute error between the two.