Supplementary Material for "A generative modeling approach for benchmarking and training shallow quantum circuits"

Marcello Benedetti, Delfina Garcia-Pintos, Oscar Perdomo,
Vicente Leyton-Ortega, Yunseong Nam, and Alejandro Perdomo-Ortiz

Comparison of cost functions

In realistic machine learning scenarios, we typically do not have access to the complete target probability distribution, nor to that obtained from the output state of a quantum circuit. Hence, we need to compare the two distributions at the level of histograms and using a finite number of samples and measurements. Here we compare three cost functions via simulations *in silico*.

First, we defined the *clipped negative log-likelihood* as

$$\mathcal{C}_{nll}(\boldsymbol{\theta}) = -\frac{1}{D}\sum_{d=1}^{D} \ln(\max(\epsilon, P_{\boldsymbol{\theta}}(\mathbf{x}^{(d)}))), \tag{S1}$$

where probabilities are estimated from samples and $\epsilon > 0$ is a small number that avoids an infinite cost when $P_{\boldsymbol{\theta}}(\mathbf{x}^{(d)}) = 0$. This may happen if for example entanglement in the circuit prevents us from measuring configuration $\mathbf{x}^{(d)}$. Moreover, when the number of variables N is large, all except few configurations will ever be measured due to the finite number of samples. Note that by re-normalizing all the probabilities after the clipping, we could interpret this variant as a Laplace additive smoothing. All the experiments in the main text are carried out using the clipped negative log-likelihood with $\epsilon = 10^{-8}$.

Second, we defined the *earth mover's distance* [S1] as

$$\mathcal{C}_{emd}(\boldsymbol{\theta}) = \min_{F} \langle d(\mathbf{x}, \mathbf{y})\rangle_F, \tag{S2}$$

where $F(\mathbf{x}, \mathbf{y})$ is a joint probability distribution such that $\sum_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}) = P_{\mathcal{D}}(\mathbf{x})$ and $\sum_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}) = P_{\boldsymbol{\theta}}(\mathbf{y})$, that is, its marginals correspond to the data and circuit distributions, respectively. Intuitively, this is the minimum cost of turning one histogram into the other where the ground metric $d(\mathbf{x}, \mathbf{y})$ specifies the cost of transporting a single unit from $\mathbf{x}$ to $\mathbf{y}$. We chose $d(\mathbf{x}, \mathbf{y})$ to be the Hamming distance between strings $\mathbf{x} \in \{-1, +1\}^N$ and $\mathbf{y} \in \{-1, +1\}^N$. Since we normalize histograms to sum up to one, the Earth Mover's Distance is equivalent to the 1-st Wasserstein distance [S2]. In our simulations, we use the `PyEMD` Python library for fast computation of the earth moving distance [S3].

Third, we defined the *moment matching* as

$$\mathcal{C}_{mm}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i}^{N}(\langle x_i\rangle_{P_{\mathcal{D}}} - \langle x_i\rangle_{P_{\boldsymbol{\theta}}})^2 + \frac{2}{N(N-1)}\sum_{i>j}^{N}(\langle x_i x_j\rangle_{P_{\mathcal{D}}} - \langle x_i x_j\rangle_{P_{\boldsymbol{\theta}}})^2, \tag{S3}$$

where the expectation values $P_{\mathcal{D}}$ and $P_{\boldsymbol{\theta}}$ are taken with respect to data and circuit distributions, respectively. This cost function can be generalized to include moments beyond the second as well as using different positive exponents for the error.

We compared the cost functions on the task of learning thermal states of size $N = 5$, with $L = 3$ layers and *all* topology. Figure S1 shows the bootstrapped median KL divergence on 25 realizations and 90% confidence interval as learning progress for 100 iterations. The fact that $\mathcal{C}_{nll}$ (red diamonds) outperforms other cost functions does not come as a surprise; minimization of the negative log-likelihood is indeed directly related to minimization of the KL divergence. However, we expect the performance of DDQCL based on this cost function to degrade quickly as the size of the problem increases. In realistic applications, the relevant probabilities in Eq. (S1), i.e. those associated with the data, are a vanishing fraction of the 2^N probabilities. Moreover, they need to be estimated from a finite number of measurements. The earth mover's distance $\mathcal{C}_{emd}$ (green pentagons) performs well, but it suffer from similar scalability issues. Fast algorithms for the computation of this distance may struggle when the number of bins in the histogram increases exponentially as in our case. However, it is reassuring to see that alternative cost functions with no relation to the KL divergence can still produce satisfactory results. Surprisingly, the moment matching $\mathcal{C}_{mm}$ (purple crosses) closely tracks the other cost functions while retaining computational efficiency. In fact, even though a large number of samples may be needed to obtain low-variance estimates for the moments, only $\mathcal{O}(N^2)$ terms are computed at each iteration. We expect this cost function to be a good heuristic for DDQCL on large systems.

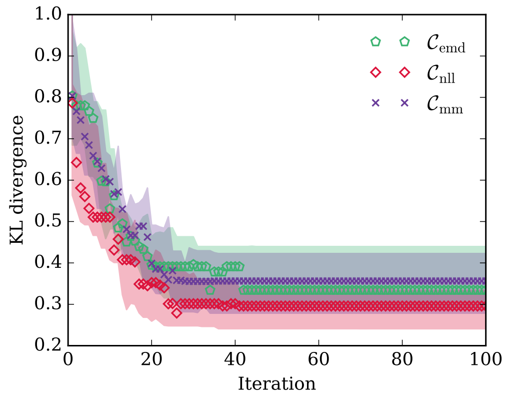

FIG. S1. Bootstrapped median KL divergence and 90% confidence interval of circuits learned by minimizing different cost functions. Both the moment matching ($\mathcal{C}_{mm}$) and earth mover's distance ($\mathcal{C}_{emd}$) closely track the clipped negative log-likelihood ($\mathcal{C}_{nll}$) used in the main text.

Approximate preparation of coherent thermal states for $N = 6$

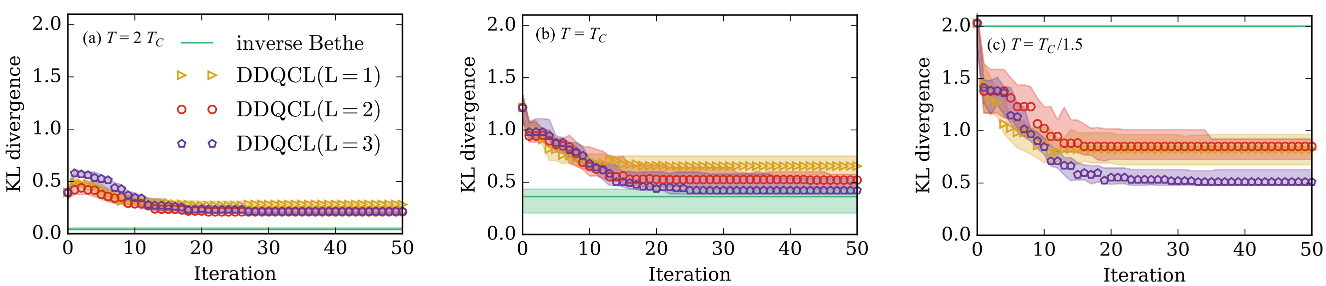

FIG. S2. DDQCL preparation of approximate coherent thermal states for the case of six qubits, different circuit depths $L \in \{1, 2, 3\}$, and different temperatures $T \in \{2T_c, T_c, T_c/1.5\}$. In these simulations, an all-to-all topology was used for the entangling gate layer ($l = 2$). We report the bootstrapped median and 90% confidence interval. For the low temperature case shown in panel (c), the inverse Bethe approximation converged only in 7 out of 25 instances. Hence, no median value was extracted. We plotted a KL divergence of 2.0 as a reference, but it is clear that for all those instances DDQCL outperformed the inverse Bethe approximation.

Details for theoretical and experimental results

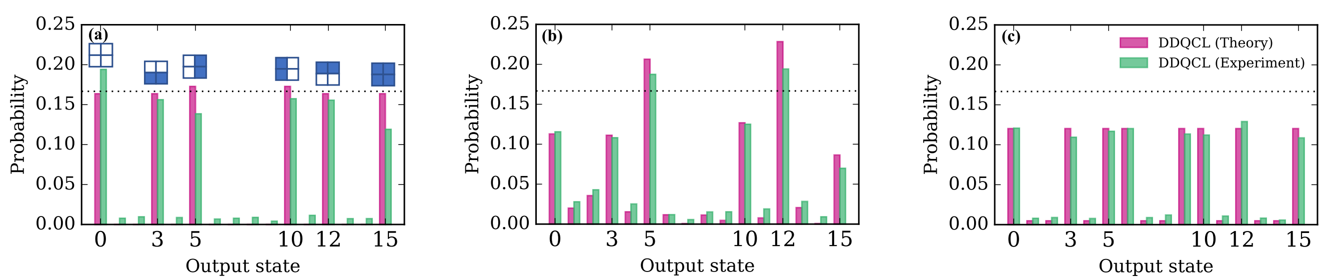

FIG. S3. Detailed comparison at the level of output states between simulation of circuits and the respective experimental implementations in the ion trap quantum computer. These are the best circuits in terms of KL divergence that were obtained by DDQCL for BAS(2, 2) under three different setting: (a) all-to-all with $L = 2$ layers, (b) star with $L = 4$, and (c) star with $L = 2$. The theoretical state obtained in (a) is close to optimal and attains an entanglement entropy averaged over all two-qubit subsets of $S_{BAS(2,2)} = 1.69989$. Circuit diagrams for (a-c) are shown in Figure S4.

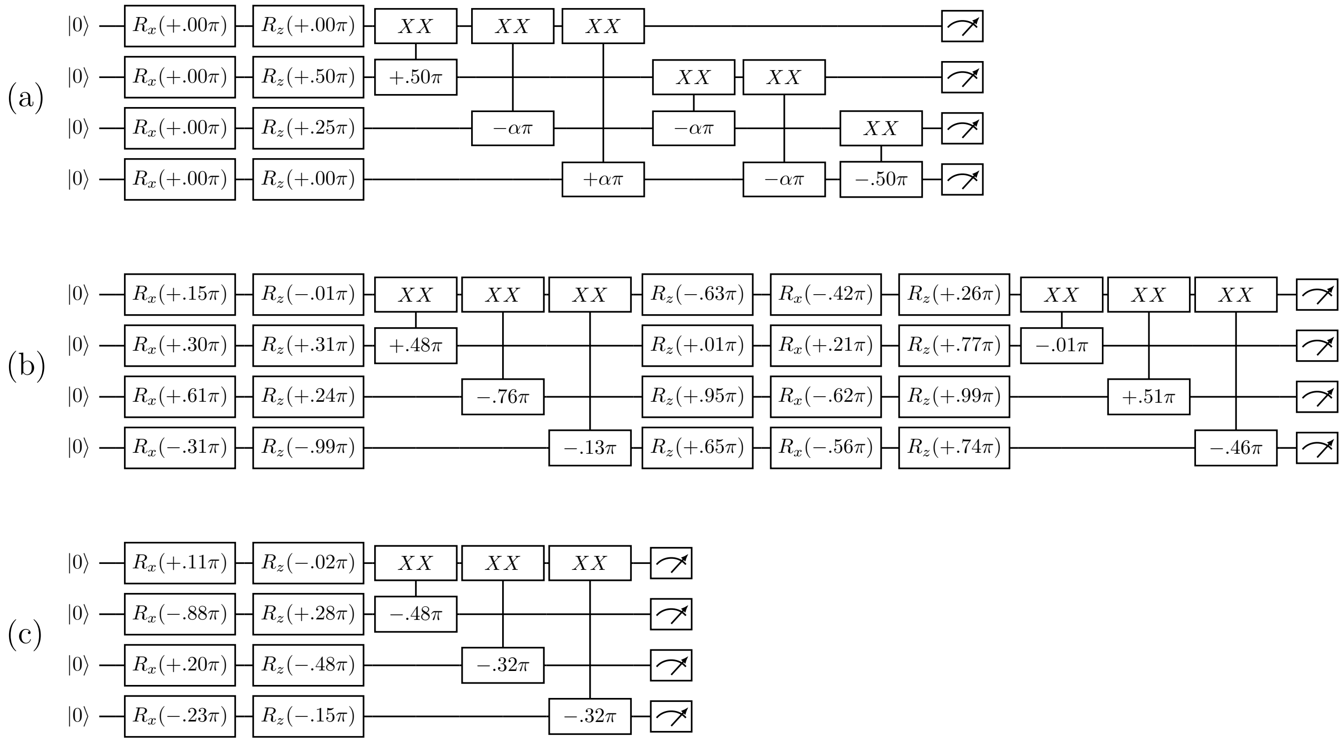

FIG. S4. Circuit diagrams for the analytical solution and for the best circuits found by DDQCL under three different qubit-to-qubit connectivity topologies. (a) the all-to-all circuit with $L = 2$ layers can achieve zero KL divergence for BAS(2, 2) by setting $\alpha = \pi^{-1} \arctan(2^{-1/2})$. All single-qubit rotations can be set to zero. DDQCL found an almost optimal solution with $\alpha = 0.2$ and two non-zero R_z rotations. These R_z gates act as the identity on the $|0000\rangle$ state. (b) the star circuit with $L = 4$. (c) the star circuit with $L = 2$. The qBAS scores for (a-c) are shown in Figure 2 in the Main Text.

Entanglement entropy of BAS(2,2)

The measure of entanglement entropy used in this work is the average von Neumann entropy over all 2-qubit subsets [S4]. Consider a 4-qubit pure state $\rho = |\psi\rangle \langle\psi|$ and label the four qubits as A, B, C, and D. Then, the entropy can be computed as

$$S_\psi = -\frac{1}{3}\Big[\mathrm{Tr}(\rho_{AB} \log_2 \rho_{AB}) + \mathrm{Tr}(\rho_{AC} \log_2 \rho_{AC}) + \mathrm{Tr}(\rho_{AD} \log_2 \rho_{AD})\Big], \tag{S4}$$

where ρ_{XY} is the reduced density matrix for the subset XY. As an example, the 4-qubit cat state has an entanglement entropy of $S_{GHZ} = 1$. Now consider a pure state encoding the uniform probability distribution over the BAS(2, 2) data set in the computational basis

$$|BAS(2,2)\rangle = \frac{1}{\sqrt{6}} \left(e^{iu_1} |0000\rangle + e^{iu_2} |0011\rangle + e^{iu_3} |0101\rangle + e^{iu_4} |1010\rangle + e^{iu_5} |1100\rangle + |1111\rangle\right). \tag{S5}$$

A direct computation shows that the entropy of this state is

$$\begin{aligned} S_{BAS(2,2)} = -\tfrac{1}{9}\Big[& \tfrac{2}{\ln(2)} \sqrt{\tfrac{\cos(u_2-u_3-u_4+u_5)+1}{2}} \tanh^{-1}\left(\sqrt{\tfrac{\cos(u_2-u_3-u_4+u_5)+1}{2}}\right) \\ & + \left(\cos\left(\tfrac{u_1-u_3-u_4}{2}\right)+1\right) \log_2\left(\tfrac{2}{3}\cos^2\left(\tfrac{u_1-u_3-u_4}{4}\right)\right) + \left(\cos\left(\tfrac{u_1-u_2-u_5}{2}\right)+1\right) \log_2\left(\tfrac{2}{3}\cos^2\left(\tfrac{u_1-u_2-u_5}{4}\right)\right) \\ & + \log_2\left(4 + 2\sqrt{2}\sqrt{\cos\left(u_2-u_3-u_4+u_5\right)+1}\right) + \log_2\left(4 - 2\sqrt{2}\sqrt{\cos\left(u_2-u_3-u_4+u_5\right)+1}\right) \\ & - \left(\cos\left(\tfrac{u_1-u_3-u_4}{2}\right)-1\right) \log_2\left(\tfrac{2}{3}\sin^2\left(\tfrac{u_1-u_3-u_4}{4}\right)\right) - \left(\cos\left(\tfrac{u_1-u_2-u_5}{2}\right)-1\right) \log_2\left(\tfrac{2}{3}\sin^2\left(\tfrac{u_1-u_2-u_5}{4}\right)\right) \\ & - \log_2(31104)\Big]. \end{aligned} \tag{S6}$$

Defining new variables $v_1 = u_2 - u_3 - u_4 + u_5$ and $v_2 = u_1 - u_3 - u_4$, the expression above reduces to

$$
\begin{aligned}
S_{BAS(2,2)} = -\tfrac{1}{9}\Big[& \tfrac{2}{\ln(2)}\sqrt{\cos^2\left(\tfrac{v_1}{2}\right)}\tanh^{-1}\left(\sqrt{\cos^2\left(\tfrac{v_1}{2}\right)}\right) \\
& + 2\cos^2\left(\tfrac{v_2}{4}\right)\log_2\left(\tfrac{2}{3}\cos^2\left(\tfrac{v_2}{4}\right)\right) + 2\cos^2\left(\tfrac{v_2-v_1}{4}\right)\log_2\left(\tfrac{2}{3}\cos^2\left(\tfrac{v_2-v_1}{4}\right)\right) \\
& + \log_2\left(4+2\sqrt{2}\sqrt{\cos\left(v_1\right)+1}\right) + \log_2\left(4-2\sqrt{2}\sqrt{\cos\left(v_1\right)+1}\right) \\
& + 2\sin^2\left(\tfrac{v_2}{4}\right)\log_2\left(\tfrac{2}{3}\sin^2\left(\tfrac{v_2}{4}\right)\right) + 2\sin^2\left(\tfrac{v_2-v_1}{4}\right)\log_2\left(\tfrac{2}{3}\sin^2\left(\tfrac{v_2-v_1}{4}\right)\right) \\
& - \log_2(31104)\Big].
\end{aligned} \tag{S7}
$$

In Figure S5 we graphically show the entropy $S_{BAS(2,2)}$ as a function of the new variables v_1 and v_2. Such a function has extrema

$$
\begin{aligned}
\min S_{BAS(2,2)} &= \frac{1}{3}\log_2\left(\frac{27}{2}\right) \approx 1.25163, \\
\max S_{BAS(2,2)} &= \frac{1}{2}\log_2(12) \approx 1.79248.
\end{aligned} \tag{S8}
$$

For the minimum value, $v_1 = v_2 = 0$, which can be obtained setting $u_1 = \cdots = u_5 = 0$. For the maximum value, $v_1 = 4\pi/3$ and $v_2 = 2\pi/3$, which can be obtained setting $u_1 = u_2 = u_3 = 0$ and $u_4 = -u_5 = 2\pi/3$. Interestingly, the maximum of $S_{BAS(2,2)}$ happens to coincide with the maximum entanglement entropy known for any 4-qubit state [S4].

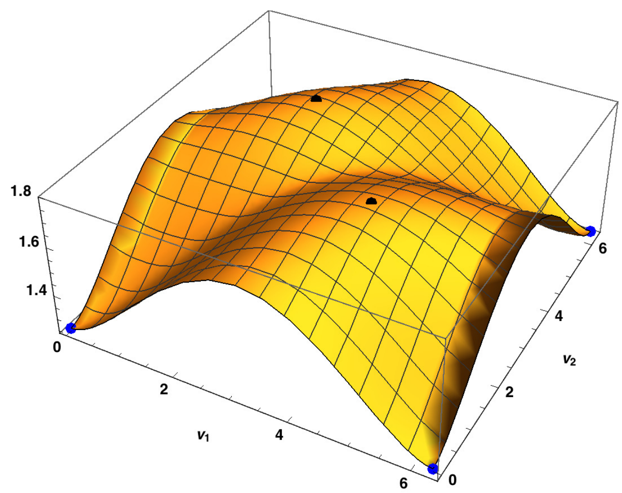

FIG. S5. Entanglement entropy $S_{BAS(2,2)}$ as a function of variables v_1 and v_2. Points in the domain represent states that encode the $\mathrm{BAS}(2,2)$ data set in the computational basis. The maximum entropy (black dots) attained is 1.79248 which coincides with the maximum entanglement entropy known for any 4-qubit state [S4].

[S1] Rubner, Y., Tomasi, C. & Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* **40**, 99–121 (2000).

[S2] Levina, E. & Bickel, P. The earth mover's distance is the mallows distance: Some insights from statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, 251–256 (IEEE, 2001).

[S3] Pele, O. & Werman, M. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, 460–467 (IEEE, 2009).

[S4] Higuchi, A. & Sudbery, A. W. How entangled can two couples get? *Physics Letters A* **273**, 213 – 217 (2000).