Featured in Physics

What Limits the Simulation of Quantum Computers?

Yiqing Zhou,[1,2] E. Miles Stoudenmire,[2] and Xavier Waintal[3]

[1]*Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[2]*Center for Computational Quantum Physics, Flatiron Institute, New York, New York 10010, USA*
[3]*Univ. Grenoble Alpes, CEA, IRIG-Pheliqs, 38054 Grenoble, France*

(Received 19 February 2020; revised 22 September 2020; accepted 5 October 2020; published 23 November 2020)

An ultimate goal of quantum computing is to perform calculations beyond the reach of any classical computer. It is therefore imperative that useful quantum computers be very difficult to simulate classically, otherwise classical computers could be used for the applications envisioned for the quantum ones. Perfect quantum computers are unarguably exponentially difficult to simulate: the classical resources required grow exponentially with the number of qubits N or the depth D of the circuit. This difficulty has triggered recent experiments on deep, random circuits that aim to demonstrate that quantum devices may already perform tasks beyond the reach of classical computing. These real quantum computing devices, however, suffer from many sources of decoherence and imprecision which limit the degree of entanglement that can actually be reached to a fraction of its theoretical maximum. They are characterized by an exponentially decaying fidelity $\mathcal{F} \sim (1-\epsilon)^{ND}$ with an error rate ϵ per operation as small as $\approx 1\%$ for current devices with several dozen qubits or even smaller for smaller devices. In this work, we provide new insight on the computing capabilities of real quantum computers by demonstrating that they can be simulated at a tiny fraction of the cost that would be needed for a perfect quantum computer. Our algorithms compress the representations of quantum wave functions using matrix product states, which are able to capture states with low to moderate entanglement very accurately. This compression introduces a finite error rate ϵ so that the algorithms closely mimic the behavior of real quantum computing devices. The computing time of our algorithm increases only linearly with N and D in sharp contrast with exact simulation algorithms. We illustrate our algorithms with simulations of random circuits for qubits connected in both one- and two-dimensional lattices. We find that ϵ can be decreased at a polynomial cost in computing power down to a minimum error ϵ_∞. Getting below ϵ_∞ requires computing resources that increase exponentially with ϵ_∞/ϵ. For a two-dimensional array of $N = 54$ qubits and a circuit with control-Z gates, error rates better than state-of-the-art devices can be obtained on a laptop in a few hours. For more complex gates such as a SWAP gate followed by a controlled rotation, the error rate increases by a factor 3 for similar computing time. Our results suggest that, despite the high fidelity reached by quantum devices, only a tiny fraction $(\sim 10^{-8})$ of the system Hilbert space is actually being exploited.

DOI: 10.1103/PhysRevX.10.041038

Subject Areas: Computational Physics, Condensed Matter Physics, Quantum Physics

I. INTRODUCTION

Operating a quantum computer is a race against the clock. The same phenomenon enabling the potential computing power of quantum computers—entanglement—is also responsible for decoherence when it occurs with unmonitored degrees of freedom. The main challenge of quantum computing is to quickly build entanglement between the qubits before imperfections or decoherence overly corrupt the quantum state. This decoherence is an intrinsic characteristic of any quantum computer and its origin and consequences must be understood thoughtfully. But in all hardware realizations, it means each operation incurs a loss of fidelity relative to the ideal target quantum state.

As different experimental platforms for quantum manipulation make rapid, impressive advances, there has been a justifiable interest in the computational capability of near-term quantum computers [1]. One of the key questions is when and how to achieve the goal of "quantum supremacy" [2], which is the crossover point where a quantum system ceases to be within reach of simulation on a classical computer. Precise circuits and fidelity metrics have been designed to meet this goal [3]. Recently, an experiment using $N = 53$ qubits and a circuit of depth

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

$D = 20$ has reached a multiqubit fidelity $\mathcal{F} = 0.002$ [4]. According to the authors of Ref. [4], such an experiment would take thousands of years to be simulated on the largest existing supercomputers. This statement was then challenged by another estimate which claims that only 2 days would be needed [5]. Such a disparity between estimates raises the question of the difficulty of simulating a quantum computer and consequently of the true computing power realized in a quantum computer.

The implicit assumption behind quantum supremacy as well as the most appealing applications of quantum computing is that a quantum computer is exponentially hard to simulate. Indeed, in recent years many techniques have been developed to simulate quantum computers, and they all have an exponential cost in some parameter. A brute force approach where one holds the full quantum state in memory as a large vector of size 2^N (N is the number of qubits) requires a computing time and memory that scales exponentially with N but linearly with the depth D of the circuit. Other approaches require a computing time that scales exponentially with the number of 2-qubit gates [6–9], with the number of non-Clifford gates [10], and/or with the number of gates that are nondiagonal in a chosen basis [11,12]. All these techniques can simulate *perfect* quantum computers. In all cases, the required computing resources are exponential so that getting beyond $N = 50$ and a depth $D = 20$ for an arbitrary circuit is extremely difficult.

In this article, we show that *real* quantum computers can be simulated at a tiny fraction of the cost that would be needed for a perfect quantum computer. To do so, we take advantage of the fact that in real quantum computers, decoherence limits the amount of entanglement that can be built into the quantum state to a fraction of what the exponentially large Hilbert space would suggest. Our algorithms use a compressed wave function representation that achieves very high accuracy for states with low to moderate entanglement. This compression introduces a finite error rate ϵ per 2-qubit gate. Hence, in this class of algorithms the limiting factor is the fidelity with which the calculation is performed while the computing time is linear in both the number of qubits N and the depth D. These algorithms "mimic" actual quantum computers both in the sense of how they scale with N and D and in the sense that the main difficulty lies in increasing the fidelity of the calculation: a small finite error ϵ is made each time a 2-qubit gate is applied to the state. Therefore, they offer a better reference point than exact simulation algorithms for assessing the computing power harvested by actual quantum chips.

Our algorithms are based on tensor networks and more precisely on matrix product states (MPS) [13]. MPS were recognized very early as an interesting parametrization of many-qubit quantum states for quantum simulations [6] and its generalizations are used in some of the most advanced quantum simulation approaches [14]. However, so far, the focus of classical simulations of quantum hardware has been building essentially exact simulations techniques and little attention has been devoted to approximate techniques. Interestingly these exact techniques can require one to go well beyond double precision calculations [15] which already hints at the link between error rate and underlying computing difficulty.

The historical success of MPS has *not* been for exact calculations but, in contrast, for the development of controlled, approximate techniques to address quantum many-body physics problems. This includes the celebrated density matrix renormalization group (DMRG) algorithm [16] which has provided precise solutions to a number of one-dimensional and quasi-one-dimensional problems, as well as time-dependent extensions [17] and generalizations to higher dimensions through projected entangled pair states (PEPS) [18] or multiscale entanglement renormalization ansatz (MERA) [19] tensor networks. At the root of these successes is the fact that MPS naturally organizes states according to the amount of entanglement entropy between different parts of the system. Hence, slightly entangled systems can be easily represented with MPS. As entanglement entropy grows, one eventually truncates the basis. The associated error can be made arbitrarily small by keeping a larger set of basis states.

In this article, we construct such an approximate technique in the context of quantum computing. Our chief result is that, for fidelities comparable to those reached experimentally, the computational requirement for simulating an imperfect quantum computer is only a tiny fraction of the requirements for a perfect one.

II. POSSIBLE STRATEGIES FOR APPROXIMATE SIMULATIONS OF QUANTUM CIRCUITS

Let us start by discussing possible strategies for simulating quantum circuits in an approximate manner. Suppose that we have partitioned the qubits into two different sets A and B with, respectively, N_A and N_B qubits ($N_A + N_B = N$). Let us consider the 2-qubit gates that connect A and B and ignore gates internal to A or B. Performing a singular value decomposition (SVD) of such a gate, it can be written as a sum of terms that act separately on A and B. This sum contains two terms for the case of usual gates (control-NOT and control-Z) and at most four terms for an arbitrary 2-qubit gate. It follows that computing the state after n of these gates amounts to keeping track of 2^n (up to 4^n) different amplitudes. These amplitudes are the discrete analog of Feynman paths and are referred to as such in the literature. For the random circuits that will be considered in this article, these 2^n amplitudes have essentially random phases. It follows that if one keeps track of just a *single* path, one reaches an overall multiqubit fidelity $\mathcal{F} = (1/2)^n$ [or $\mathcal{F} = (1/4)^n$ in the worst situation]. This very simple strategy could be used to simulate an arbitrary large number of qubits with low fidelity per gate in a computing time $\sim n$.

However, if one wants to keep a fixed fidelity per gate f defined as $\mathcal{F} = f^n$, in analogy with real quantum computers, the number of paths N_{path} that must be tracked during the simulation is $N_{\text{path}} = (2f)^n$, and hence increases exponentially with n. Such a strategy has been used in Ref. [4] to validate the experimental results reported there.

We now seek algorithms where a constant fidelity f can be obtained at a constant computing cost per gate, independent of the total number of gates n. One starts by writing a general state for the bipartite system as

$$|\Psi\rangle = \sum_{a,b} \Psi_{ab} |a\rangle_A |b\rangle_B, \tag{1}$$

where the states $|a\rangle$ ($|b\rangle$) form an orthonormal basis of A (B). Performing a singular value decomposition,

$$\Psi_{ab} = \sum_{\mu} U_{a\mu} S_\mu V_{\mu b}, \tag{2}$$

one can define an orthonormal basis,

$$|\mu\rangle_A = \sum_a U_{\mu a} |a\rangle_A \tag{3}$$

(with similar notation for the B subsystem), and arrive at the usual Schmidt decomposition of $|\Psi\rangle$:

$$|\Psi\rangle = \sum_{\mu} S_\mu |\mu\rangle_A |\mu\rangle_B, \tag{4}$$

in terms of a finite number of singular values S_μ. States with only one nonzero singular value $S_0 = 1$ are simple, unentangled product states. A measure of the number of significant singular values needed to describe the state to high accuracy is given by the entanglement entropy,

$$S = -\text{Tr}\rho_A \log \rho_A = -\text{Tr}\rho_B \log \rho_B = -\sum_{\mu} S_\mu^2 \log S_\mu^2, \tag{5}$$

where ρ_A (ρ_B) is the reduced density matrix for the subsystem A (B). The general strategy of DMRG-like algorithms is to keep only a finite number χ of the singular values. After a 2-qubit gate that connects A and B, one performs a SVD decomposition of Ψ_{ab} and truncates the state by keeping only the χ largest singular values. When $\chi \gg e^S$, this procedure is essentially exact. As the entanglement increases, this procedure leads to a certain fidelity per gate $f < 1$ that can be controlled by increasing the parameter χ. Of interest to the present article is the typical value of f that can be reached in a reasonable computing time.

III. NOISY ALGORITHM IN ONE DIMENSION

Above we motivated the truncated SVD of a bipartite wave function as an approximation strategy that works well for wave functions with only a moderate amount of entanglement. A natural generalization of this strategy to the N-qubit case is to use matrix product states, which can be viewed as a simultaneous Schmidt decomposition of the wave function across N different partitions [6] or equivalently a sequence of compatible SVD factorizations of the wave function, grouping qubits $1, 2, \ldots, j$ and $j+1, \ldots, N$ and performing an approximate SVD of the resulting matrix [13].

A. MPS representation of the state

We first consider a one-dimensional network of N qubits where 2-qubit gates can only be applied directly between nearest neighbors. (Within this connectivity, gates acting on other non-neighboring qubits are still possible at the cost of using $\sim N$ SWAP operations to bring the qubits onto neighboring sites.) We define our MPS state in terms of N tensors $M(n)$ as

$$\begin{aligned} |\Psi\rangle &= \sum_x \Psi_x |x\rangle \\ &= \sum_{i_1 \ldots i_N} \sum_{\mu_1 \ldots \mu_{N-1}} M(1)^{i_1}_{\mu_1} M(2)^{i_2}_{\mu_1 \mu_2} M(3)^{i_3}_{\mu_2 \mu_3} \ldots \\ &\quad M(N)^{i_N}_{\mu_{N-1}} |i_1 i_2 i_3 \ldots i_N\rangle, \end{aligned} \tag{6}$$

where the "physical" indices $i_n \in \{0, 1\}$ span the 2^N-dimensional Hilbert space while the bond (or virtual) indices $\mu_n \in \{1, \ldots, \chi_n\}$ control the maximum degree of entanglement allowed by the MPS. $|x\rangle$ is a shorthand for $|i_1 i_2 \ldots i_N\rangle$. If the χ_n are allowed to grow exponentially large as a function of N, then the MPS form of the wave function becomes exact and can represent any wave function [13]. In contrast, we will enforce $\chi_n \leq \chi$ in what follows so that the resulting MPS represents an approximation of the true wave function. The parameter χ controls the error rate made by our algorithm as well as the computational and memory costs required to run it. As we see below, applying a 2-qubit gate takes $\sim \chi^3$ operations and the overall memory footprint is $N\chi^2$. A sketch of the MPS structure is shown in Fig. 1(b).

To be acceptable, our algorithm must provide the same features that a real quantum computer would provide. Applying a 1-qubit gate U on qubit n can be done exactly and without increasing any of the χ_n: it simply amounts to updating the corresponding tensor $M(n) \to M'(n)$:

$$M'(n)^{i'_n}_{\mu_{n-1}\mu_n}(n) = \sum_{i_n} U_{i'_n i_n} M(n)^{i_n}_{\mu_{n-1}\mu_n}, \tag{7}$$

as shown in Fig. 2(a). Calculating the overlap between different MPS states or calculating individual wave function amplitudes $\langle i_1 i_2 \ldots i_{N-1} i_N | \Psi \rangle$ can be done with contraction

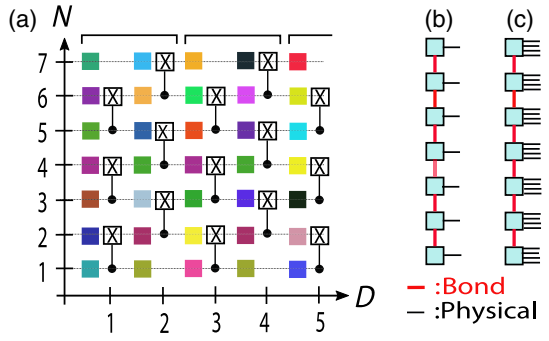

FIG. 1. (a) Sketch of the quantum circuit with N qubits. The colored squares indicate arbitrary 1-qubit gates while the dots connected to a cross indicate a 2-qubit gate such as control-NOT or control-Z. The depth D counts the number of 2-qubit gates performed in the sequence. (b) Structure of the matrix product states (MPS) for 1D circuits. Red lines indicate bond (or virtual) indices while thin black lines correspond to physical indices. (c) MPS structure for quasi-one-dimensional structures.

algorithms which, for MPS, can be done exactly in $\sim N\chi^3$ operations (see, e.g., Ref. [13] for a detailed description of standard MPS algorithms). It follows that one can also sample from the distribution $|\langle i_1 i_2 \ldots i_{N-1} i_N|\Psi\rangle|^2$ within the same complexity. Quantum measurements (sampling of a given qubit followed by its projection) can also be done efficiently in a straightforward manner [20].

To perform a 2-qubit gate U between qubit n and qubit $n+1$, one first transforms the MPS into an "orthogonal form" centered around the qubits of interest [13]. More precisely, we perform a series of QR factorizations from left to right to bring the tensors on the left of tensor n into a

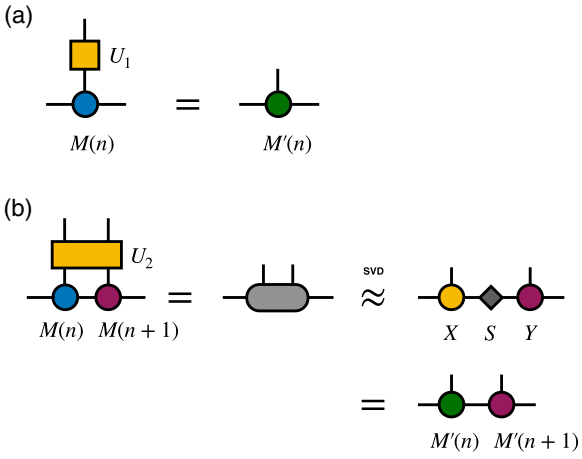

FIG. 2. (a) Applying a single qubit gate to a MPS can be done without approximation by multiplying the gate by a single MPS tensor. (b) To apply a 2-qubit gate to qubits n and $n+1$, one contracts the corresponding tensors together, then applies the gate. To restore the MPS form, the resulting tensor is decomposed with a SVD truncated to keep the largest χ singular values, and the matrix of singular values is multiplied into one of the unitary factors X or Y.

"left-orthogonal" form. We perform another series of LQ factorizations (QR on the transpose) from right to left to bring the tensors on the right of tensor $n+1$ into a "right-orthogonal" form. Bringing the MPS into this form is crucial for the accuracy of truncations of the MPS and we have observed that without it the error per gate of the algorithm would be about 2 times larger. The steps to apply the gate are then shown in Fig. 2(b). One first forms the 2-qubit tensor:

$$T^{i_n i_{n+1}}_{\mu_{n-1}\mu_{n+1}} = \sum_{\mu_n} M(n)^{i_n}_{\mu_{n-1}\mu_n} M(n+1)^{i_{n+1}}_{\mu_n\mu_{n+1}}. \tag{8}$$

Then one applies the 2-qubit gate U and obtains

$$(T')^{i'_n i'_{n+1}}_{\mu_{n-1}\mu_{n+1}} = \sum_{i_n i_{n+1}} U_{i'_n i'_{n+1}, i_n i_{n+1}} T^{i_n i_{n+1}}_{\mu_{n-1}\mu_{n+1}}. \tag{9}$$

In a last stage, considering the tensor T' as a matrix with indices spanned by (i'_n, μ_{n-1}) and (i'_{n+1}, μ_{n+1}), one performs a singular value decomposition and writes

$$(T')^{i'_n i'_{n+1}}_{\mu_{n-1}\mu_{n+1}} = \sum_{\mu_n} X^{i'_n}_{\mu_{n-1}\mu_n} S_{\mu_n} Y^{i'_{n+1}}_{\mu_n\mu_{n+1}}, \tag{10}$$

where the tensors X and Y are formed of orthogonal vectors while the vector S_μ contains the singular values of T'. Here S_μ has up to 2χ components (irrespective of the nature of the 2-qubit gate) so that exact algorithms imply a doubling of χ after each application of a 2-qubit gate. In the spirit of DMRG-like algorithms, we truncate S_μ and keep only its χ largest components to obtain S'_μ. The new MPS tensors are then simply given by

$$M'(n)^{i_n}_{\mu_{n-1}\mu_n} = X^{i_n}_{\mu_{n-1}\mu_n} S'_{\mu_n}, \tag{11}$$

$$M'(n+1)^{i_{n+1}}_{\mu_n\mu_{n+1}} = Y^{i_{n+1}}_{\mu_n\mu_{n+1}}, \tag{12}$$

which completes the algorithm. Overall, the cost of applying a 2-qubit gate is dominated by the SVD step which scales as χ^3. We emphasize that such an algorithm can do anything that a quantum computer does, but the reverse statement is not true: in the MPS approach, one holds the full wave function in memory, which provides much more information than can be obtained from samples of the wave function. For instance, one can compute bipartite entanglement entropy of a MPS, and it is straightforward to calculate quantities such as observables or correlation functions without any statistical errors. The MPS format also satisfies the sample and query access criteria needed for quantum inspired dequantizing algorithms [21].

B. Random quantum circuit

Figure 1(a) shows the quantum circuit used in our numerical experiments. It consists of alternating layers of 1-qubit and 2-qubit gates. This circuit has been designed following the proposal of Ref. [3] in order to create strongly entangled states in as few operations as possible. It is believed to be one of the most difficult circuits to simulate on a classical computer since its many-qubit quantum state is extremely sensitive to modification of any of the gates. The 1-qubit gates U_n represented as colored squares in Fig. 1(a) are chosen randomly such as to remove any structure or symmetry from the many-qubit state. A gate U_n is a rotation $U_n = \exp(-i\theta_n \vec{\sigma}.\vec{m}_n)$ of angle θ_n around a unit vector $\vec{m}_n = (\sin\alpha_n \cos\phi_n, \sin\alpha_n \sin\phi_n, \cos\alpha_n)$ ($\vec{\sigma}$ is the vector of Pauli matrices). We take the angles θ_n, α_n, and ϕ_n to be uniformly distributed. Note that the resulting matrix U_n is *not* distributed according to the Haar distribution of $U(2)$. We have, however, also experimented with other choices of distribution of the 1-qubit gates (including the Haar measure) and found that it did not affect our results except for small variations of the obtained fidelity. While the U_n are random, the actual sequence used is carefully recorded for comparison with, e.g., exact calculations. We call the number of 2-qubit gate layers applied the depth D of the circuit, focusing on the number of 2-qubit gate layers because those are the only source of imperfection in our calculations. In real quantum computers, 2-qubit gates also dominate the errors over 1-qubit gates in terms of fidelity. However, real quantum computers also have other sources of error (decoherence, unknown couplings between qubits, leakage to noncomputational states, etc.) not present in the algorithm. After a depth $D \sim N$, the state obtained with the circuit of Fig. 1(a) is totally scrambled and well described by a Porter-Thomas distribution. This is illustrated in Fig. 3, where the cumulative distribution of $p_x = |\langle x|\Psi\rangle|^2$ is compared to the Porter-Thomas form for various maximum MPS bond dimensions (main panel) and for various depths using exact calculations (inset). One indeed observes that the distribution quickly approaches the chaotic Porter-Thomas distribution as one increases the bond dimension χ.

C. Effective 2-qubit gate fidelity

Let us introduce the main quantity of interest for this study, the effective 2-qubit fidelity f_n. The effective 2-qubit fidelity f_n is the computational analog to the fidelity reported experimentally for 2-qubit gates. $f_n = 1$ for a perfect calculation, but the truncation of the MPS will induce $0 < f_n < 1$.

Let us call $|\Psi_T(n)\rangle$ the MPS state after a sequence of n individual 2-qubit gates [$n \approx (N-1)D/2$ for the circuit of Fig. 1(a)]. Up to irrelevant 1-qubit gates, $|\Psi_T(n)\rangle$ is obtained by applying one control-Z gate C_Z onto $|\Psi_T(n-1)\rangle$ followed by the truncation operation

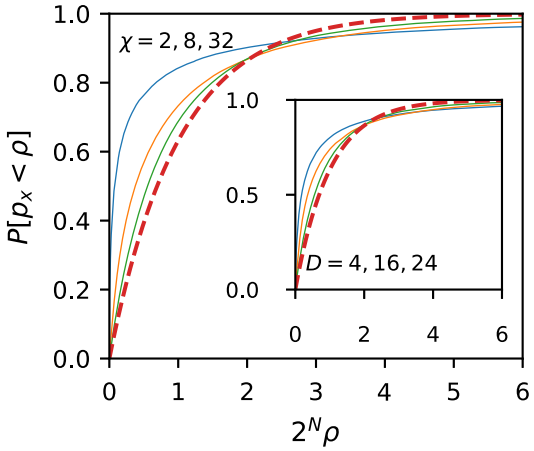

FIG. 3. Cumulative distribution $P(p_x < \rho)$, where $p_x = |\langle x|\Psi\rangle|^2$ for $N = 15$. The dashed line corresponds to the Porter-Thomas distribution, $P_{\mathrm{PT}}(\rho) = 1 - (1-\rho)^{2^N-1}$. Main panel: MPS simulations with truncation, $D = 24$, and various MPS truncation levels $\chi = 2$ (blue line), 8 (orange line), and 32 (green line). Inset: exact results (i.e., simulations without truncations) for $D = 4$ (blue line), 16 (orange line), and 24 (green line). A single realization of the circuit has been used for each distribution.

which introduces a finite error. We define the effective fidelity f_n as

$$f_n = |\langle \Psi_T(n)|C_Z|\Psi_T(n-1)\rangle|^2, \tag{13}$$

and the corresponding error rate ϵ_n as

$$\epsilon_n = 1 - f_n. \tag{14}$$

f_n can be calculated using the contraction algorithm in $N\chi^3$ operations. However, when the MPS is in canonical form, f_n is simply obtained without any additional calculations as

$$f_n = \left(\sum_{\mu=1}^{\chi} S_\mu^2\right) \Big/ \left(\sum_{\mu=1}^{2\chi} S_\mu^2\right), \tag{15}$$

where we recall that 2χ is the maximum possible number of nonzero singular values of the tensor T' in Eq. (10). The denominator above is always equal to one for a state which is normalized before it is acted on by a 2-qubit gate. We have explicitly checked the equivalence between the two algorithms.

A typical simulation is shown in Fig. 4 for the circuit with the control-Z gate. At small depth $D < 2\log_2\chi$, the simulation is exact and $f_n = 1$. Above this threshold, one starts to truncate the MPS after each 2-qubit gate. We observe a transient regime where f_n decreases after which f_n quickly saturates at a constant value, here around 0.988. The first thing to note in Fig. 4 is that these simulations are many orders of magnitude easier than an equivalent *perfect* calculation: simulating the exact state for $N = 60$ and

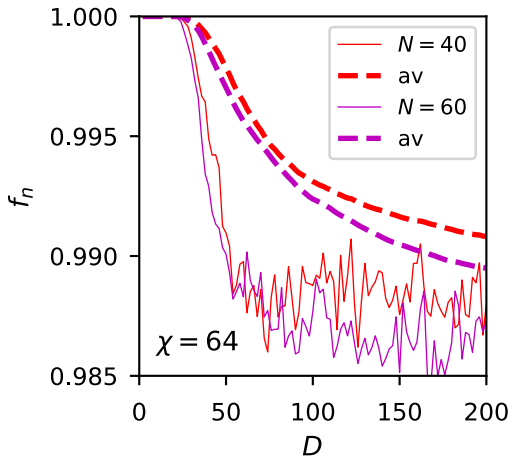

FIG. 4. Effective 2-qubit gate fidelity f_n as a function of the depth D of the circuit for $\chi = 64$ and the control-Z gate for $N = 40$ (red) and $N = 60$ (magenta). The thin lines correspond to the geometric average of f_n over one full sequence, i.e., all the 2-qubit gates performed between depth $D-2$ and depth D ($N-1$ 2-qubit gates). The thick dashed lines correspond to f_{av}, the geometric average of f_n over all 2-qubit gates since the beginning of the circuit up to depth D. A single realization of the circuit has been used for each curve.

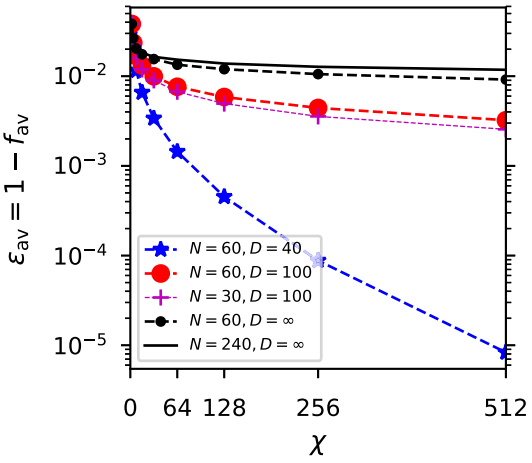

FIG. 5. Geometric average of the residual error per gate $\epsilon_{\mathrm{av}} = 1 - f_{\mathrm{av}}$ as a function of the bond dimension χ. The average is performed over the entire circuit except for the black curves ($D = \infty$) where the average is restricted to the regime where the fidelity has reached its asymptotic value ($100 \leq D \leq 200$); see Fig. 4. For the largest system $N = 240$, we have also excluded the gates on the edges of the system in our calculation as they have by construction perfect fidelity. The fluctuations of the average fidelity with different circuits are smaller than the size of the symbols.

$D = 200$ would be out of reach even with thousands of years of computing time on the largest existing supercomputer. Yet here, these simulations of a *noisy* quantum computer have been performed on a laptop. The averaged fidelity for a modest $\chi = 64$ is better than 99%, which already corresponds to qubits of very good quality. This is rather remarkable since the percentage of the Hilbert space spanned by the MPS ansatz is only a very tiny fraction $\sim 10^{-13}\%$ of the whole Hilbert space. After the transient regime, f_n is, up to some fluctuations, independent of both D and N. The second statement is true up to small $1/N$ corrections. These corrections arise from the fact that the fidelity associated with gates applied on the edge of the system [i.e., associated to matrices $M(i)$ with $i < 2\log_2\chi$ or $N - i < 2\log_2\chi$] is always equal to unity since the entanglement entropy associated to the subsystem of qubits $i < a$ is bounded by $S \leq a \log 2$.

Our main goal is to understand how the residual error $\epsilon_n = 1 - f_n$ decreases as one increases the bond dimension χ. As χ approaches $\chi = 2^{N/2}$, one must have $\epsilon_n \to 0$. This is because reshaping the wave function as a $2^{N/2} \times 2^{N/2}$ matrix implies a maximum rank of $2^{N/2}$ for any factorization of the wave function; thus a MPS with $\chi = 2^{N/2}$ remains exact. However, here we are interested in the regime $\chi \ll 2^{N/2}$ which remains accessible to simulations. Figure 5 shows how the residual error $\epsilon_n = 1 - f_n$ decreases with increasing the bond dimension. The main finding of Fig. 5 is that the residual error per gate at large depth D and number of particle N eventually saturates at a finite value, in this case around $\epsilon_\infty \approx 10^{-2}$. In other words, this algorithm can simulate any 1D quantum computer that has a 2-qubit gate fidelity smaller than $f_\infty = 99\%$ at a *linear* cost in both N and D. As the depth or number of qubits is reduced, the average fidelity increases. The black cross in Fig. 5 corresponds to a calculation where only the last part of the circuit has been taken into account in the calculation of the average fidelity; i.e., the average is performed for $D > 100$ where the system has already entered its stationary regime. Note that in that regime, there remains a small logarithmic decrease of the error: as χ increases, a number $\propto \log_2\chi$ of gates close to the edges of the system become exact, as discussed above. The black line in Fig. 5 corresponds to calculations made in a larger system of $N = 240$ qubits where we have restricted the calculation of the fidelity to the gates for qubits in the center of the system (i.e., away from the edges where the fidelity is perfect) as well as removed the small depth regime (only gates for $100 \leq D \leq 200$ are taken into account). For this case, we observe a clear saturation of the error rate to a finite value ϵ_∞. As we shall see, decreasing the error rate beyond ϵ_∞ requires an exponential effort.

Figure 6 shows the dependence of the fidelity on the position n where the gate is applied. The gates applied on the edges, i.e., between qubit n and $n+1$ such that $2^{\min(n,N-n)} < \chi$, are always exact ($f = 1$). As χ increases, more and more gates on the edge become exact until eventually all gates become exact when χ becomes exponentially large ($\chi = 2^{N/2}$). Away from the edges, we observe a clear plateau at f_∞ in the large N and χ limit. Numerically, we cannot exclude that this plateau has a very

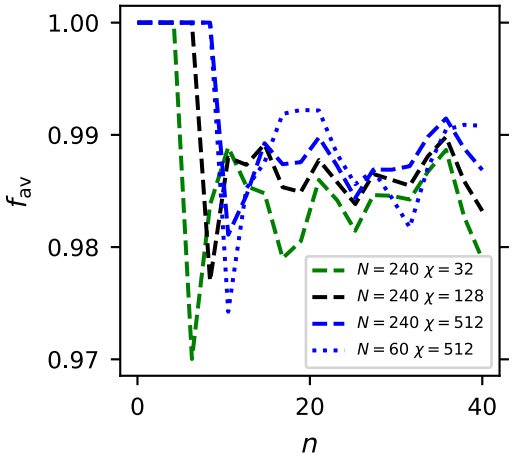

FIG. 6. Geometric average of the fidelity per gate f_{av} as a function of the gate position n. Only an enlargement close to one edge of the MPS is shown. The average is restricted to the regime where the fidelity has reached its asymptotic value ($100 \leq D \leq 200$); see Fig. 4. A single realization of the circuit has been used, hence the fluctuations. These fluctuations become very small upon further averaging on n as done in Fig. 5.

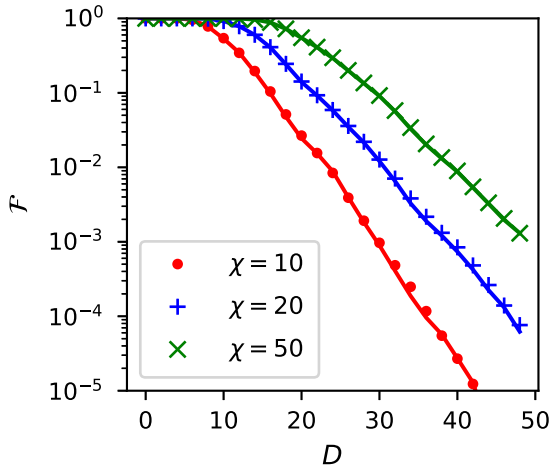

FIG. 7. Fidelity $\mathcal{F}$ versus depth D for $N = 20$ and various values of $\chi = 10, 20, 50$. The symbols correspond to a direct calculation of $\mathcal{F}$ obtained by comparing with an exact calculation (i.e., without truncation). The lines correspond to the right-hand side of Eq. (17).

slow $1/\chi^a$ with $a < 0.2$ or logarithmic decrease. However, further evidence for the existence of a true plateau are given by the distribution of singular values discussed in Sec. V.

IV. LINKS BETWEEN 2-QUBIT AND MULTIQUBIT FIDELITY

Before investigating the origin of ϵ_∞, we make a short detour to discuss how the effective 2-qubit fidelity f_n is related to the actual N-qubit fidelity $\mathcal{F}$ of the state and is related to practical estimates of the fidelity that can be measured experimentally.

A. Multiqubit fidelity

Let us call $|\Psi_P(n)\rangle$ the exact perfect state after n 2-qubit gates—meaning it is never truncated or otherwise approximated at any stage of its evolution by the circuit—while $|\Psi_T(n)\rangle$ is the truncated MPS state (P stands for perfect and T for truncated). The N-qubit fidelity $\mathcal{F}$ is defined as

$$\mathcal{F}(n) = |\langle\Psi_P(n)|\Psi_T(n)\rangle|^2. \tag{16}$$

The fidelity $\mathcal{F}$ is a direct measure of how reliable our truncated state is. As the errors accumulate, it is natural to expect that the fidelities f_n are multiplicative:

$$\mathcal{F}(n) \approx \prod_{i=1}^{n} f_i. \tag{17}$$

Equation (17) is indeed a very accurate approximation. An analytical argument is given below. The validity of Eq. (17) can also be shown by numerical simulations. Figure 7 shows the fidelity versus D for $N = 20$ particles obtained in two independent ways. The symbols correspond to a direct calculation of $\mathcal{F}$ while the lines correspond to the right-hand side of Eq. (17). We find an almost perfect match in all the regimes that we have studied. Equation (17) is a very useful result: it relates a property of the perfect state (left-hand side) to a property solely defined in terms of the MPS (right-hand side). It allows us to easily estimate the fidelity in regimes where we do not have access to the exact state anymore. When f_n has reached its stationary value f_∞, Eq. (17) simplifies into

$$\mathcal{F}(n) \approx (f_\infty)^n \sim (f_\infty)^{ND/2}. \tag{18}$$

In an actual experiment, one cannot measure the f_n, but rather one has access to an estimate of $\mathcal{F}(n)$ (see the section below). To compare the accuracy of the simulations with the capabilities of actual quantum chips, we therefore define the average 2-qubit fidelity f_{av} after n 2-qubit gates,

$$f_{av} = \left(\prod_{i=1}^{n} f_i\right)^{1/n} \approx \mathcal{F}(D)^{2/ND}, \tag{19}$$

where the second equality is specific to the quantum circuit studied here.

Derivation of Eq. (17).—Let us define a full basis of orthogonal states $|\alpha\rangle$ such that state $|1\rangle \equiv |\Psi_T(n-1)\rangle$ is our truncated state and we complement state $|1\rangle$ with an arbitrary basis. Writing $|\Psi_P(n-1)\rangle$ in that basis as $|\Psi_P(n-1)\rangle = \sum_{\alpha=1}^{2^N} p_\alpha|\alpha\rangle$, we have $p_1 = \sqrt{\mathcal{F}(n-1)}$. Similarly, we write $|\Psi_T(n)\rangle = \sum_{i=1}^{2^N} t_\alpha C_Z|\alpha\rangle$, with $t_1 = \sqrt{f_n}$. From these definitions, the fact that C_Z is unitary and that $|\Psi_P(n)\rangle = C_Z|\Psi_P(n-1)\rangle$, we have

$$\mathcal{F}(n) = \left[\sum_{\alpha=1}^{2^N} p_\alpha t_\alpha\right]^2 = \left[\sqrt{\mathcal{F}(n-1)f_n} + \sum_{\alpha=2}^{2^N} p_\alpha t_\alpha\right]^2. \tag{20}$$

As the fidelity goes down, the p_α and t_α become increasingly decorrelated, in particular in sign. Assuming random signs between the p_α and the t_α and using that $p_\alpha \sim 1/\sqrt{2^N}$, we find that the second term in the above equation is at most of order $1/\sqrt{2^N}$ and is therefore negligible. Equation (17) follows directly.

We end this section by proving a weaker but exact bound for shallow circuits without the above assumption.

The Schwartz inequality implies that

$$\left(\sum_{\alpha=2}^{2^N} p_\alpha t_\alpha\right)^2 \le \sum_{\alpha=2}^{2^N} p_\alpha^2 \sum_{\alpha=2}^{2^N} t_\alpha^2 \le \epsilon_n, \tag{21}$$

from which we obtain

$$|\sqrt{\mathcal{F}(n)} - \sqrt{f_n \mathcal{F}(n-1)}| \le \sqrt{\epsilon_n}. \tag{22}$$

The Eq. (22) bound is exact, but saturating this bound in practice implies that all the terms $p_\alpha t_\alpha$ interfere constructively, which is not realized in actual circuits. Equation (22) implies that

$$\begin{aligned}\mathcal{F}(n) &\ge \left[\sqrt{f_n \mathcal{F}(n-1)} - \sqrt{\epsilon_n}\right]^2 \\ &\ge \mathcal{F}(n-1) - 2\sqrt{\epsilon_n},\end{aligned} \tag{23}$$

from which one can prove that

$$\mathcal{F}(n) \ge 1 - 2\sum_{i=1}^{n} \sqrt{\epsilon_i}\,. \tag{24}$$

The exact statement Eq. (24) can be useful for small depth circuits where the actual decrease of the fidelity $\mathcal{F}(n)$ is indeed linear with n, before one enters into the true exponential regime.

B. Other fidelity metrics

So far we have used the overlap $\mathcal{F}$ between the exact state $|\Psi_P\rangle$ and our approximate state $|\Psi_T\rangle$ as our metric for the fidelity of the calculation. It is a natural metric as it measures the probability for the approximate state to be in the exact state. It is bounded $0 \le \mathcal{F} \le 1$ and is nicely related to the probabilities per gate f_n through the formulas of the preceding section.

However, $\mathcal{F}$ cannot be directly measured experimentally, so that other fidelity metrics must be designed. Indeed, in an actual quantum computer, the only existing output are samples of bit strings $x = i_1 i_2 \ldots i_N$ distributed according to $|\langle x|\Psi_T\rangle|^2$. A natural metric is the logarithmic cross entropy defined as

$$\mathcal{C} = -\sum_x |\langle x|\Psi_T\rangle|^2 \log |\langle x|\Psi_P\rangle|^2. \tag{25}$$

Logarithmic cross entropy is a standard tool of machine learning and has several interesting properties. First, it is measurable through sampling as

$$\mathcal{C} = -\lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} \log |\langle x_m|\Psi_P\rangle|^2, \tag{26}$$

where the x_m are the output of the quantum computer when the experiment is repeated M times. Second, the logarithmic cross entropy between two distributions $|\langle x|\Psi_T\rangle|^2$ and $|\langle x|\Psi_P\rangle|^2$ is maximum when the two distributions are identical. Hence it is a genuine measure of the likelihood of the two distributions. Logarithmic cross entropy was proposed in Ref. [3] as a fidelity metric. Note, however, that the logarithmic cross entropy is not a symmetric function of the two distributions. In particular, it is strongly affected by particular configurations x, where $|\langle x|\Psi_P\rangle|^2$ is very low but $|\langle x|\Psi_T\rangle|^2$ is not.

Logarithmic cross entropy was eventually abandoned by the Google team and replaced [4] by the linear cross entropy benchmarking (XEB) defined as

$$\mathcal{B} = -1 + 2^N \sum_x |\langle x|\Psi_T\rangle|^2 |\langle x|\Psi_P\rangle|^2. \tag{27}$$

XEB is also sampleable and is symmetric with respect to the two distributions. When the approximate state is the uniform distribution, the XEB metric vanishes, $\mathcal{B} = 0$, indicating a total lack of fidelity. However, when the approximate state is actually exact, the value of the XEB metric can be arbitrary. When the approximate state is exact *and* distributed according to the Porter-Thomas distribution (which happens in our circuits after a few cycles), then the XEB metric gets a well-defined $\mathcal{B} = 1$ value. The XEB metric is not in general a good measure of the likelihood between two distributions: for a given perfect state; it is maximum when the approximate state is sharply peaked around the values of x where the perfect state is maximum. In our circuit the initial value of XEB is exponentially high $\mathcal{B} = 2^N - 1$ and quickly decreases as the distribution approaches the Porter-Thomas one. Calling D^* the depth after which XEB has reached unity (ideally D^* would the depth after which $|\langle x|\Psi_P\rangle|^2$ corresponds to Porter-Thomas distribution), we find empirically that

$$\mathcal{F}_n \approx \mathcal{F}(D^*)\mathcal{B}_n. \tag{28}$$

Equation (28) could be used to estimate the actual fidelity $\mathcal{F}$ from XEB measurements.

Figure 8 shows an example of calculations contrasting the fidelity $\mathcal{F}$ with the XEB metric (see also Ref. [22]). Here we have used no truncation but added some noise on

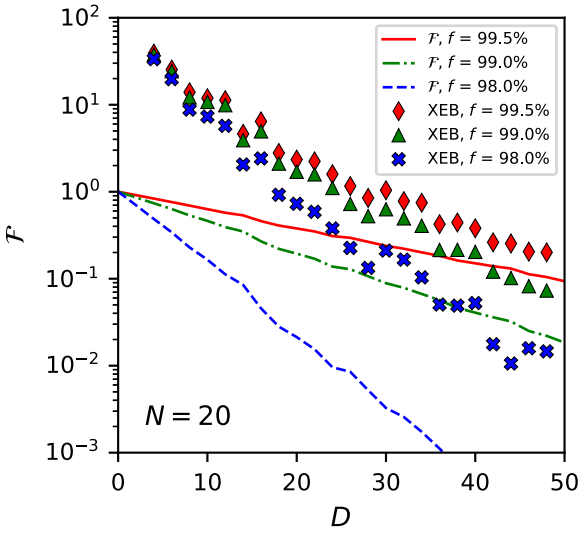

FIG. 8. Comparison between the fidelity $\mathcal{F}$ (lines) and the XEB metric $\mathcal{B}$ (markers) as a function of depth D. Different colors label different levels of noise on the 2-qubit gates, respectively, $f = 99.5\%$ (red), $f = 99\%$ (green), and $f = 98\%$ (blue). The calculations were performed for the 1D random circuit with $N = 20$ qubits.

the 2-qubit gate so as to induce a finite fidelity per gate f. We find that both $\mathcal{F}$ and XEB decay exponentially with consistent decay rates. However, the large difference of the initial values at $D = 0$ leads to a shift of the fidelity, which is significantly lower than the XEB curve. This shift increases as the fidelity is lowered. For a typical experimental value $f = 99\%$, we find that XEB overestimates the fidelity $\mathcal{F}$ by about a factor 10 in our simulations.

V. RANDOM TENSOR THEORY OF ϵ_∞

We now turn back to the discussion of the asymptotic value f_∞ reached by the 2-qubit gate fidelity in our calculations. The first remark of importance is that f_∞ is a property associated with a single tensor of the full MPS state: if we apply a gate between qubit i and qubit $i+1$, only the associated T' tensor defined in Eq. (10) comes into play. Since the whole goal of our quantum circuit is to scramble the wave function as efficiently as possible, a natural hypothesis is that the tensors $M(i)$ and $M(i+1)$ become eventually well described by totally random tensors. In this section we explore this possibility and calculate the properties of the associated tensor T' as well as the corresponding 2-qubit gate fidelity f_{GTE}. We find that the distribution of singular values of T' obtained from the random ensemble closely matches what we observe in the MPS state.

In the spirit of random matrix theory [23,24], we introduce the Gaussian tensor ensemble (GTE) where a tensor $M^i_{\mu\nu}$ is supposed to be totally random. The GTE can be thought of as a "worst-case scenario" where the quantum circuit is so chaotic that the tensors are left with no structure. In the GTE, the tensors M are distributed according to

$$P[M^i_{\mu\nu}] \propto \exp\left[-\frac{1}{2}\sum_{\mu\nu i}|M^i_{\mu\nu}|^2\right], \tag{29}$$

where the sum over ν spans $1\ldots\chi$, the sum over i spans 0,1, and the sum over μ spans $1\ldots\beta\chi$. In the remainder of this section, we restrict ourselves to $\beta = 1$, which corresponds to the tensors of Eq. (6). We shall have an example of $\beta = 2$ for the grouped-qubit algorithm we discuss in Sec. VI. From two such tensors, we apply a 2-qubit gate following Eqs. (8)–(12) constructing the associated tensor T and T' and the SVD of T'. From the $2\beta\chi$ singular values S_μ of T', we can obtain the associated fidelity f_{GTE} through Eq. (15).

Figure 9 studies the distribution of the singular values S_μ for tensor T' obtained from the GTE. The singular values are sorted in order of decreasing magnitude and plotted as a function of the index $\mu = 1, \ldots, 2\chi$. Plotting χS^2_μ as a function of μ/χ, we observe that all the different values of χ collapse onto a single curve. In other words, we find that there is some function $g(x)$ such that

$$S^2_\mu = \frac{1}{\chi} g\left(\frac{\mu}{\chi}\right). \tag{30}$$

This scaling is already valid for rather small values of χ. This observation can probably be put on firm mathematical grounds—it is consistent with the usual scaling of the semicircular law of the so-called Gaussian unitary

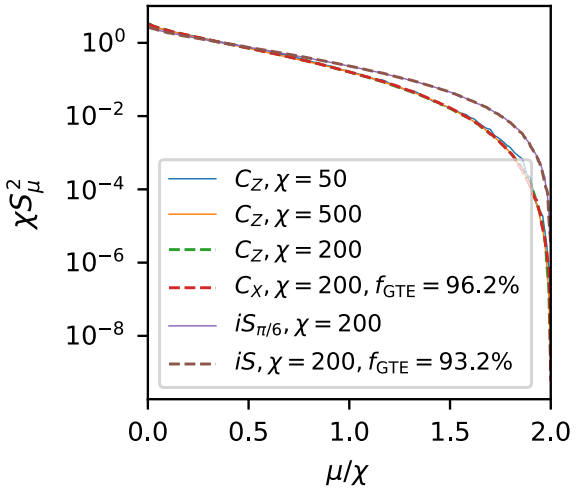

FIG. 9. Squared singular values S^2_μ of the matrix T' obtained from the GTE ensemble. We find a perfect scaling of the form $S^2_\mu = g(\mu/\chi)/\chi$, where μ is the index of the μth singular value. The two bundles of curves correspond respectively to the C_X,C_Z gates (two nonzero eigenvalues) and the $iS_{\pi/6}/iS$ gates (four nonzero eigenvalues). Within one bundle, the different curves are indistinguishable.

ensemble—but for the moment it is merely an empirical statement made from numerical evidence. It follows from this scaling that f_{GTE} very quickly converges to

$$f_{\mathrm{GTE}} = \frac{\int_0^1 dx g(x)}{\int_0^{2\beta} dx g(x)}. \tag{31}$$

In other words, one finds a finite value of the fidelity that is independent of χ. The resulting f_{GTE} depends, on the other hand, on the 2-qubit gate used. Control-Z (C_Z) and control-NOT (C_X) are equivalent (they are related to each other through a change of basis of the second qubit) and correspond to $f_{\mathrm{GTE}} = 96.2\%$. Gates like the iSWAP gate (iS) or iSWAP followed by a $\pi/6$ rotation over the z axis ($iS_{\pi/6}$, close to what is used in Ref. [4]) have four different singular values, which roughly doubles the error with respect to C_Z ($f_{\mathrm{GTE}} = 93.2\%$).

Figure 10 shows how the distribution of the singular values in the GTE compares to the one obtained in the MPS simulation. We find a close agreement between GTE and the MPS simulations when looking at the T' tensor for a gate in the center of the system and at large depth. The agreement is not perfect, however, and we observe that the asymptotic fidelity of MPS simulations is always better than the one found in GTE:

$$f_\infty \geq f_{\mathrm{GTE}}. \tag{32}$$

Equation (32) is a numerical observation that reflects the fact that a random structureless MPS is a worst-case scenario for our truncation algorithm. To try to understand why the inequality in Eq. (32) is not fully saturated, we plot in Fig. 10 the distribution of the singular value of the initial tensor M (dotted line). After truncation, the distribution of the singular values of M' are given by the dashed line restricted to $0 \leq \mu/\chi \leq 1$ (up to a small shift due to the normalization of the state). These two distributions differ very significantly. In order to saturate the bound of Eq. (32) we would need extra steps to scramble the distribution of M' back to the distribution of M (i.e., go from the dashed line to the dotted line). However, since in our protocol only a single 1-qubit gate separates one truncation from the next one, we find that it is not sufficiently chaotic and therefore we never reach the "worst-case scenario" of the GTE.

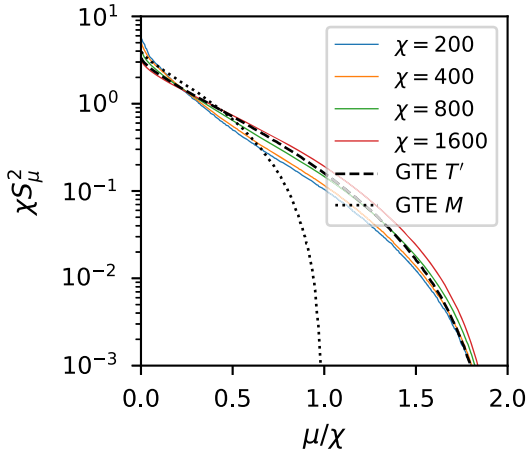

FIG. 10. Squared singular values S_μ^2 of the matrix T' obtained from the MPS simulations of $N = 30$ qubits and a depth of $D = 60$ for various values of χ. The singular values correspond to a gate C_X performed in the middle of the system. Dotted line: squared singular values of the M matrix in the GTE. Dashed line: squared singular values of T' in the GTE.

To summarize, f_{GTE} can be thought of as a lower bound for the fidelity found in the simulations for large enough χ (typically $\chi \geq 300$ in practice) and large enough depth. Getting beyond the asymptotic value requires algorithms that have an exponential cost. In the following section we describe possible strategies.

VI. ALGORITHMS FOR GETTING BEYOND ϵ_∞

The algorithm discussed above can also be used for 2D arrays, since any 2-qubit gates between distant qubits can always be written as a combination of gates on neighboring qubits using SWAP gates. However, this is inefficient and leads to a decrease of the effective f as the transverse dimension of the 2D array increases. Another limitation of the above algorithm is that one cannot efficiently simulate systems that have a fidelity above f_∞.

There are multiple strategies that could be used to go beyond the above algorithm. In particular, recent progress in the algorithms for contracting tensor networks, such as Ref. [9], could be interesting candidates in 2D. Below, we follow a very simple strategy where we keep using MPS states, but group the qubits so that each tensor now represents several qubits. The idea is to perform several 2-qubit gates per truncation, thereby lowering the error per gate. We show that this strategy works in practice up to unexpectedly large fidelities at moderate computational cost. We surmise that it may be pursued to arbitrary small error rate at an exponential computational cost, but this point remains to be further investigated. The grouped MPS algorithms used below are actually quasi-one-dimensional algorithms: the computational cost scales linearly with the numbers of columns but exponentially with the number of qubits per column.

A. Grouped MPS state and extraction algorithm

We now consider the MPS structure sketched in Fig. 1(c), where each tensor addresses several qubits. We now have $P \leq N$ tensors $M(n)$ each addressing N_n qubits with $\sum_{n=1}^{P} N_n = N$. The tensors $M(1)$ and $M(P)$ possess $N_n + 1$ indices while the others possess $N_n + 2$ indices:

$$M(n)_{\mu\nu}^{i_1 i_2 \ldots i_{N_n}}. \tag{33}$$

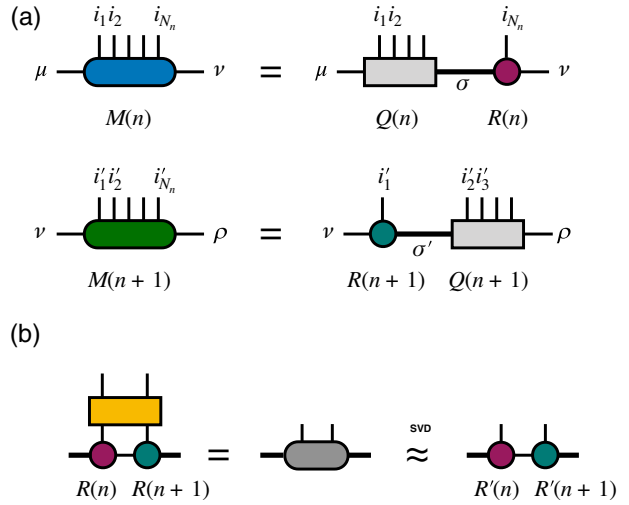

FIG. 11. Main steps for applying a gate which acts across two grouped MPS tensors, as described in Eqs. (34) and (35). In (a) the grouped MPS tensors $M(n)$ and $M(n+1)$ are exactly factorized using QR decompositions, such that the $R(n)$ and $R(n+1)$ tensors carry the qubit indices acted on by the gate and the newly introduced indices σ and σ' range over 2χ values. In (b) the gate acts on the product of $R(n)$ and $R(n+1)$, and the resulting tensor is factorized using a SVD truncated to χ singular values. Finally, to update the MPS (not shown), one computes the new tensors $M'(n) = Q(n)R'(n)$ and $M'(n+1) = R'(n+1)Q(n+1)$, which diagrammatically looks like step (a) but in reverse.

The number of elements of these tensors is $\chi^2 2^{N_n}$ so that the computing time now increases exponentially with the number of qubits per tensor. On the other hand, the 2-qubit gates that are performed inside a given tensor $M(n)$ are now handled exactly, so that the average fidelity of a circuit increases.

To perform a 2-qubit gate between neighboring tensors $M(n)$ and $M(n+1)$, one proceeds in three steps. The first two are shown diagrammatically in Fig. 11. In the first step, one performs a QR decomposition of the two tensors to "extract" smaller tensors corresponding to the involved qubits. Assuming (without loss of generality) that the 2-qubit gate involves qubit N_n of tensor $M(n)$ and qubit 1 of tensor $M(n+1)$, one decomposes $M(n)$ as

$$M(n)_{\mu\nu}^{i_1 i_2 \ldots i_{N_n}} = \sum_{\sigma=1}^{2\chi} Q(n)_{\mu,\sigma}^{i_1 i_2 \ldots i_{N_n-1}} R(n)_{\sigma,\nu}^{i_{N_n}}, \qquad (34)$$

where the "vectors" of $Q(n)$ indexed by σ are orthonormal. The important point here is that the index σ takes only 2χ values. Similarly, we write

$$M(n+1)_{\nu\rho}^{i'_1 i'_2 \ldots i'_{N_{n+1}}} = \sum_{\sigma=1}^{2\chi} R(n+1)_{\nu,\sigma'}^{i'_1} Q(n+1)_{\sigma',\rho}^{i'_2 \ldots i'_{N_{n+1}}}. \qquad (35)$$

The second step follows Eqs. (8)–(12) of the algorithm of Sec. III with the replacement $M(n) \to R(n)$ and $M(n+1) \to R(n+1)$, and is shown for the present case in Fig. 11(b). In the last step, the new tensors $M'(n)$ and $M'(n+1)$ are obtained by contracting $Q(n)$ with $R'(n)$ and $R'(n+1)$ with $Q(n+1)$.

The main difference between the algorithm of Sec. III and the grouped MPS algorithm is that the resulting tensor T' of Eq. (10) now has 4χ singular values instead of 2χ. As a result, upon truncation to keep only χ singular values, we anticipate that the fidelity per gate will be smaller than in the 1D case. However, as we shall see, this decrease will be more than compensated by the gain of having perfect gates within one tensor. In the terminology of random tensors, the grouped MPS algorithm corresponds to $\beta = 2$. For the C_Z gate, the GTE fidelity drops from $f_{\mathrm{GTE}}(\beta = 1) = 96.2\%$ down to $f_{\mathrm{GTE}}(\beta = 2) = 87.4\%$.

B. Application to a two-dimensional circuit

We now show the results of simulations performed on a 2D circuit. To put the results into the perspective of what can be achieved experimentally, we choose a circuit very close to the one used by the Google team in their "supremacy" experiment [4]. We consider a 2D grid of 54 qubits as shown in Fig. 12(a). The circuit is shown in Fig. 12(b) and alternates 1-qubit gates applied to each qubit (same distribution as in the 1D case) with 2-qubit gates (control-Z) applied on different pairs of qubits according to the color shown. Except for the choices of 1- and 2-qubit gates, and the number of qubits (53 versus 54), the setup is identical to the "supremacy sequence" of the Google experiment [4]. In Ref. [4] a XEB fidelity $\mathcal{B} = 0.002$ was reached after a depth $D = 20$ corresponding to a total of 430 2-qubit gates. Ignoring the difference between XEB and the fidelity $\mathcal{F}$, this translates into $\epsilon_{\mathrm{av}} = 1.4\%$, which we shall use as our reference value to evaluate the performance of the grouped MPS algorithm.

Figure 12(c) shows various strategies for grouping the qubits. The $[1^{12}]$ grouping corresponds to 12 tensors that contain one column of qubit each (i.e., alternatively 5 and 4 qubits). The [6, 6] grouping is the most expensive computationally with two tensors of 27 qubits each. Note that the tensors on the edges are less computationally costly than the middle ones, since they only have one bond index. The results of the simulations are shown in Fig. 13 for a depth of $D = 20$. While the error rate is significantly larger than in the 1D case, we find that it can be brought down to less than 1.4% (which corresponds to a global fidelity of $\mathcal{F} = 0.002$) on a single core computer. The computing times of the data points of Fig. 13 range from a few seconds to less than 48 hours for the most expensive points on a nonparallel code (single core calculation). We find that the grouping strategy is effective, but not as efficient as the maximum gain that one could expect: even though some of the gates become perfect upon grouping, we observe a decrease of

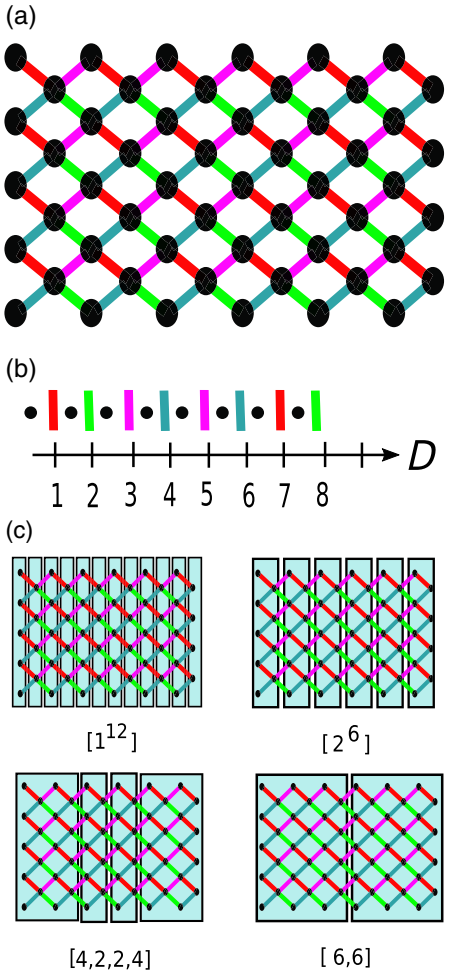

FIG. 12. (a) Sketch of the quantum circuit with 54 qubits in a 2D grid. The qubits are represented by the black dots while the 2-qubit gates by the color links. (b) The circuit alternates 1-qubit gates (black dots) with 2-qubit gates (here the control-Z gate). The depth D counts the number of 2-qubit gates per qubit. (c) Different grouping strategies for the group MPS algorithm. $[1^{12}]$ corresponds to a grouping in 12 blocks counting 1 column each; $[4,2,2,4]$ corresponds to a grouping in 4 blocks counting, respectively, 4, 2, 2, and 4 columns.

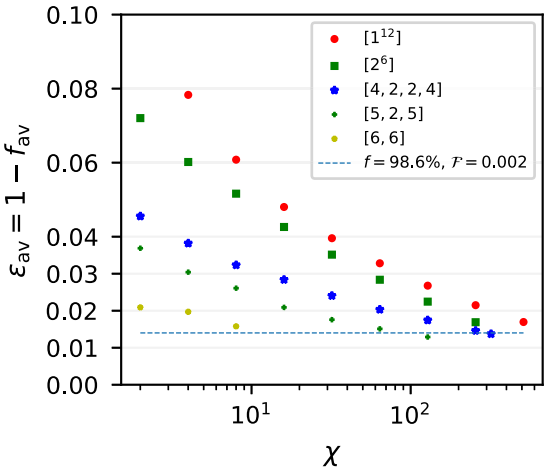

FIG. 13. Residual error per gate $\epsilon_{av} = 1 - f_{av}$ as a function of the bond dimension χ for the 2D circuit of Fig. 12 for a depth $D = 20$. The different curves correspond to different groupings. The horizontal dashed line corresponds to the error rate associated with a global fidelity $\mathcal{F} = 0.002$.

the fidelity for the noisy gates which reduces the overall gain. For $\chi = 320$ and the $[4,2,2,4]$ partition where the final fidelity is slightly better than $\mathcal{F} = 0.002$ (see Fig. 13), the memory footprint of the calculation is 4.5 GB of memory, which represents only $1.5 \times 10^{-6}\%$ of the size of the total Hilbert space spanned by the 2^{54} qubits.

C. Split-and-merge algorithm for more complex gates

We end this article with results in a configuration that closely matches the one of Ref. [4]. The 1-qubit gates are chosen at random between $\sqrt{X}$, $\sqrt{Y}$, and $\sqrt{W}$ while the 2-qubit gate iS_θ is a combination of iSWAP followed by a controlled rotation along the z axis:

$$iS_\theta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 \\ 0 & -i & 0 & 0 \\ 0 & 0 & 0 & e^{-i\theta} \end{pmatrix}. \tag{36}$$

This gate has four different singular values and is therefore expected to produce more entanglement than the C_Z gate. The link between number of singular values and the actual growth of entanglement is not totally straightforward, however. Indeed, the pure iSWAP gate has four different singular values ± 1 and $\pm i$, yet as it preserves the structure of product states, it is trivial to simulate with perfect fidelity. In what follows, we use $\theta = 1$, which is nontrivial to simulate.

The algorithm of the previous section behaves rather poorly for the iS_θ gate. For instance, for $\chi = 128$, and the $[4,2,2,4]$ grouping, the 2-qubit gate fidelity drops from $f \approx 98\%$ (C_Z) to $f \approx 92\%$ (iS_θ). However, a simple modification of the algorithm allows one to recover a much higher fidelity, $f \approx 95\%$.

To study iS_θ, we therefore switch to a "split-and-merge" strategy: instead of "extracting" qubits one by one to perform 2-qubit gates as in Sec. VI A, we extract one full column of qubits at a time. In the split-and-merge strategy, we use two different groupings of the qubits, for instance, switching between the $[4,2,2,4]$ grouping and the $[5,2,5]$ grouping (hereafter referred to as the $[4,2,2,4] \leftrightarrow [5,2,5]$ grouping strategy). Switching from one grouping to another induces truncation errors. However, once the switching has been done, many 2-qubit gates can be performed exactly. A schematic of the split-and-merge strategy is shown in Fig. 14 for the $[4,2,2,4] \leftrightarrow [5,2,5]$ case.

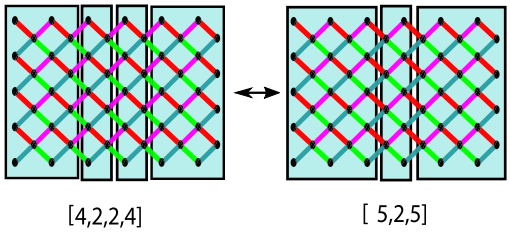

FIG. 14. Schematic of the split-and-merge algorithm for the $[4,2,2,4] \leftrightarrow [5,2,5]$. The 2-qubit gates shown in red and dark green are performed in the $[4,2,2,4]$ configuration and one switches to the $[5,2,5]$ to perform the light green and purple gates.

Figure 15 shows our numerical results for ϵ_{av} versus χ. The curves are very similar to those obtained for C_Z at similar computational cost, but with an error rate roughly 3 times larger than with C_Z.

To conclude this section, we have shown that for the control-Z gate a simple grouping strategy allows one to reach the same fidelity as the Google experiment [4] in a matter of hours on a single core computer (i.e., $f_{\text{av}} \geq 98.6\%$). For the more challenging iS_θ gate, this fidelity drops down to 95% for similar computing time.

A natural question that arises is whether these algorithms may be used to defeat the claim of quantum supremacy put forward in Ref. [4], i.e., raise the fidelity from 95% to $> 98\%$. We have not be able to do so on a single core implementation. However, the split-and-merge algorithm is to a large extent trivially parallelizable since most tensor operations contain "spectator" indices whose different values can be fixed, and the resulting tensor "slices" dispatched to different computing cores or nodes.

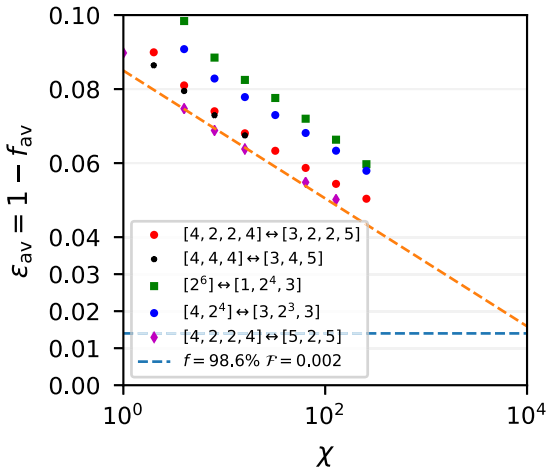

FIG. 15. Residual error per gate $\epsilon_{\text{av}} = 1 - f_{\text{av}}$ as a function of the bond dimension χ for the iS_θ gate for a 2D circuit with $N = 54$ qubits and a depth $D = 20$. The different curves correspond to different groupings. The horizontal dashed line corresponds to the error rate associated with a global fidelity $\mathcal{F} = 0.002$. The orange line is just a guide to the eye.

Extrapolations from our results suggest that such a parallel implementation should be able to reach fidelities in the 98%–99% range with a few hundred cores and a few terabytes of memory. However, such a calculation has not be attempted at the moment. Let us note, in any case, that not too much emphasis should be put on quantum supremacy by itself. It is not because a task is difficult to simulate that it provides a useful output. Also, there is no question that quantum many-body problems are extremely difficult to simulate. The insight that we get from the present work is an estimate of the relation between the accuracy reached in the quantum state and the underlying amount of entanglement that could potentially be exploited.

VII. DISCUSSION

In this work, we have discussed a practical algorithm that allows us to simulate a quantum computer in a time which grows linearly with the number of qubits N and the depth D at the cost of having a finite fidelity f per 2-qubit operation. Hence, although we do not aim at describing the actual errors and decoherence mechanisms present in real quantum computers, our algorithm provides quantum states of the same quality provided that the effective fidelity f is as high as the experimental one. The fidelity f can be increased at a polynomial cost up to a finite value f_∞; increasing it further has an exponential cost in the fidelity. Our main observation is that fidelities of the order of 99%, which are typical fidelities found in state-of-the-art experiments, can be reproduced at a moderate computational cost.

Is a fidelity of 99% large or small? From an experimental physics perspective, it is certainly quite an achievement to keep several dozen qubits at this level of fidelity. From a quantum information and classical algorithms point of view, a question is, what is the level of entanglement—hence the actual fraction of the Hilbert space that can truly been accessed—associated with this level of fidelity? Our MPS ansatz can provide an estimate (or at least an upper bound for one may come up with better algorithms) for this fraction. Since the MPS ansatz only spans a very tiny fraction of the overall Hilbert space, it follows that the computational power associated with fidelities in the 99% range is much more limited than the full size 2^N of the Hilbert space would suggest. We conclude that increasing the computational power of a quantum computer will primarily require increasing the fidelity with which the different operations are performed [25]. Increasing the number of qubits will remain ineffective until better fidelities have been reached.

A second factor of primary importance is qubit connectivity: Long-range connections mean that entanglement over much larger distances can be built before decoherence steps in. Architectures that try to improve the connectivity with, e.g., quantum buses [26] could be a very effective way to make the system harder to simulate, hence increase its potential computing power. We have indeed observed that

2D systems are much more difficult to simulate than 1D ones. Part of this difficulty is intrinsic to the increased connectivity of 2D systems. Another part is due to our MPS ansatz being not well adapted to 2D geometries. Generalization of MPS to 2D, such as PEPS, would probably be more efficient. As the PEPS representation has recently been adapted to time evolution algorithms [27], such a generalization should be rather direct.

As a side comment, our approach could also be used to get lower bounds for quantum error correction (QEC) schemes [28]. Suppose that for a certain connectivity, one has an algorithm that can reach a fidelity f in polynomial time in N and D. Then, it is reasonable to expect that any QEC code has a threshold $p > f$. If it were not the case, one could build a logical quantum computer with a classical one at a polynomial cost by simply simulating the QEC protocols on the classical computer. In this respect, extending our approach to a truly 2D algorithm (beyond the quasi-1D one discussed in this article) would be particularly interesting. Indeed, 2D surface codes have a particularly low threshold $p \approx 99\%$. How close to $f = 99\%$ one can get at a polynomial cost in 2D is currently an open question. Note that the above reasoning supposes that an algorithm that can simulate random circuits can also simulate any other circuit with a similar or a larger fidelity. While this assumption is commonly made, it remains to be rigorously proven.

Finally, it would be interesting to perform a similar study, but of how well MPS of practical sizes can approximate circuits designed for specific and useful tasks such as the Shor or Grover algorithms. It would be interesting to see if random circuits are indeed harder to simulate than more structured ones, as often implicitly assumed. Goals could include estimating minimum fidelities needed to perform these tasks with a high success probability and understanding crossovers where useful quantum algorithms begin to offer advantages over classical approaches.

ACKNOWLEDGMENTS

X. W. and Y. Z. thank the Flatiron CCQ where this work was initiated. X. W. acknowledges funding from the French ANR QCONTROL and the EU FET open UltraFastNano. Numerical results involving MPS were obtained using the ITensor library [29]. The Flatiron Institute is a division of the Simons Foundation. We thank Thomas Ayral for interesting discussions.

[1] J. Preskill, *Quantum Computing in the NISQ Era and Beyond*, Quantum **2**, 79 (2018).

[2] J. Preskill, *Quantum Computing and the Entanglement Frontier*, arXiv:1203.5813.

[3] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, *Characterizing Quantum Supremacy in Near-Term Devices*, Nat. Phys. **14**, 595 (2018).

[4] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell *et al.*, *Quantum Supremacy Using a Programmable Superconducting Processor*, Nature (London) **574**, 505 (2019).

[5] E. Pednault, J. A. Gunnels, G. Nannicini, L. Horesh, and R. Wisnieff, *Leveraging Secondary Storage to Simulate Deep 54-Qubit Sycamore Circuits*, arXiv:1910.09534.

[6] G. Vidal, *Efficient Classical Simulation of Slightly Entangled Quantum Computations*, Phys. Rev. Lett. **91**, 147902 (2003).

[7] J. Chen, F. Zhang, C. Huang, M. Newman, and Y. Shi, *Classical Simulation of Intermediate-Size Quantum Circuits*, arXiv:1805.01450.

[8] C. Guo, Y. Liu, M. Xiong, S. Xue, X. Fu, A. Huang, X. Qiang, P. Xu, J. Liu, S. Zheng, H.-L. Huang, M. Deng, D. Poletti, W.-S. Bao, and J. Wu, *General-Purpose Quantum Circuit Simulator with Projected Entangled-Pair States and the Quantum Supremacy Frontier*, Phys. Rev. Lett. **123**, 190501 (2019).

[9] F. Pan, P. Zhou, S. Li, and P. Zhang, *Contracting Arbitrary Tensor Networks: General Approximate Algorithm and Applications in Graphical Models and Quantum Circuit Simulations*, Phys. Rev. Lett. **125**, 060503 (2020).

[10] S. Aaronson and D. Gottesman, *Improved Simulation of Stabilizer Circuits*, Phys. Rev. A **70**, 052328 (2004).

[11] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, and H. Neven, *Simulation of Low-Depth Quantum Circuits as Complex Undirected Graphical Models*, arXiv:1712.05384.

[12] B. Jónsson, B. Bauer, and G. Carleo, *Neural-Network States for the Classical Simulation of Quantum Computing*, arXiv:1808.05232.

[13] U. Schollwck, *January 2011 Special Issue, The Density-Matrix Renormalization Group in the Age of Matrix Product States*, Ann. Phys. (Amsterdam) **326**, 96 (2011).

[14] I. L. Markov and Y. Shi, *Simulating Quantum Computation by Contracting Tensor etworks*, SIAM J. Comput. **38**, 963 (2008).

[15] A. SaiToh, *A Multiprecision C++ Library for Matrix-Product-State Simulation of Quantum Computing: Evaluation of Numerical Errors*, J. Phys. Conf. Ser. **454**, 012064 (2013).

[16] S. R. White, *Density Matrix Formulation for Quantum Renormalization Groups*, Phys. Rev. Lett. **69**, 2863 (1992).

[17] S. Paeckel, T. Khler, A. Swoboda, S. R. Manmana, U. Schollwck, and C. Hubig, *Time-Evolution Methods for Matrix-Product States*, Ann. Phys. (Amsterdam) **411**, 167998 (2019).

[18] F. Verstraete, V. Murg, and J. I. Cirac, *Matrix Product States, Projected Entangled Pair States, and Variational Renormalization Group Methods for Quantum Spin Systems*, Adv. Phys. **57**, 143 (2008).

[19] G. Vidal, *Entanglement Renormalization*, Phys. Rev. Lett. **99**, 220405 (2007).

[20] A. J. Ferris and G. Vidal, *Perfect Sampling with Unitary Tensor Networks*, Phys. Rev. B **85**, 165146 (2012).

[21] N.-H. Chia, A. Gilyén, T. Li, H.-H. Lin, E. Tang, and C. Wang, *Sampling-Based Sublinear Low-Rank Matrix Arithmetic Framework for Dequantizing Quantum Machine Learning*, arXiv:1910.06151.

[22] A. Dang, *Distributed Matrix Product State Simulations of Large-Scale Quantum Circuits* Master's thesis, The Univeristy of Melbourne, 2017 (https://doi.org/11343/239081).

[23] M. L. Mehta, *Random Matrices*, Pure and Applied Mathematics, 3rd ed. (Elsevier, London, 2004).

[24] C. W. J. Beenakker, *Random-Matrix Theory of Quantum Transport*, Rev. Mod. Phys. **69**, 731 (1997).

[25] X. Waintal, *What Determines the Ultimate Precision of a Quantum Computer*, Phys. Rev. A **99**, 042318 (2019).

[26] C. Bauerle, D. C. Glattli, T. Meunier, F. Portier, P. Roche, P. Roulleau, S. Takada, and X. Waintal, *Coherent Control of Single Electrons: A Review of Current Progress*, Rep. Prog. Phys. **81**, 056503 (2018).

[27] M. P. Zaletel and F. Pollmann, *Isometric Tensor Network States in Two Dimensions*, Phys. Rev. Lett. **124**, 037201 (2020).

[28] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*, 10th ed. (Cambridge University Press, New York, 2011).

[29] M. Fishman, S. R. White, and E. M. Stoudenmire, *The ITensor Software Library for Tensor Network Calculations*, arXiv:2007.14822.