The Cost of Entanglement Renormalization on a Fault-Tolerant Quantum Computer

Joshua Job,[1] Isaac H. Kim,[2] Eric Johnston,[2] and Steve Adachi[1]

[1] *Lockheed Martin, Sunnyvale, CA, 94089*
[2] *PsiQuantum, Palo Alto, CA, 94304*

(Dated: April 18, 2024)

We perform a detailed resource estimate for the prospect of using deep entanglement renormalization ansatz (DMERA) on a fault-tolerant quantum computer, focusing on the regime in which the target system is large. For probing a relatively large system size (64×64), we observe up to an order of magnitude reduction in the number of qubits, compared to the approaches based on quantum phase estimation (QPE). We discuss two complementary strategies to measure the energy. The first approach is based on a random sampling of the local terms of the Hamiltonian, requiring $\mathcal{O}(1/\epsilon^2)$ invocations of quantum circuits, each of which have depth of at most $\mathcal{O}(\log N)$, where ϵ is the relative precision in the energy and N is the system size. The second approach is based on a coherent estimation of the expectation value of observables averaged over space, which achieves the Heisenberg scaling while incurring only a logarithmic cost in the system size. For estimating the energy per site of ϵ, $\mathcal{O}\left(\frac{\log N}{\epsilon}\right)$ T gates and $\mathcal{O}\left(\log N\right)$ qubits suffice. The constant factor of the leading contribution is shown to be determined by the depth of the DMERA circuit, the gates used in the ansatz, and the periodicity of the circuit. We also derive tight bounds on the variance of the energy gradient, assuming the gates are random Pauli rotations.

arXiv:2404.10050v2 [quant-ph] 17 Apr 2024

I. INTRODUCTION

One of the most impactful applications of quantum computers is expected to be the simulation of realistic materials that appear in nature [1]. This idea, which goes back to the seminal work of Feynman [2] and Lloyd [3], has recently gained much interest with the rapid advancement in quantum computing technology.

Many of the applications considered in the literature had been focused on simulation of a small molecule associated with the nitrogen fixation process, known as "FeMoco" in the literature. Reiher et al. [4] was the first to perform a rigorous resource estimate for this system. Since then, improvements were made in the efficiency of this algorithm by adopting the block-encoding approach of Low and Chuang [5] and by using cleverly designed ways to block-encode the Hamiltonian [6–8].

More recently, quantum algorithms for simulating material properties of a solid state system have been studied [9–13]. The primary challenge in this direction is that the number of orbitals one needs to consider is much larger than that required for simulating small molecules. While there are many motivating examples, let us consider the case of multiferroic materials [14, 15] as a concrete example. These materials, which exhibit a coupling between their magnetic (magnetization) and electric (polarization) properties, have been a subject of intense experimental and theoretical interest over the past few decades. A prototypical example of a multiferroic is an epitaxially grown $BiFeO_3$ thin film [16]. These materials consist of domains, exhibiting either electric or antiferromagnetic order [17] and the bulk properties of the material depend on the sizes of these domains. The typical side of the domain consists of 100×100 unit cells if not larger [18]. Therefore, to accurately simulate realistic properties of this material, one would need to incorporate at least tens of thousands of orbitals, at least two orders of magnitude larger than the number required for small molecules [4, 6–8].

These new class of algorithms [9–13] — known as first quantization methods — have better asymptotic scaling for the number of qubits and the T-gate/Toffoli count compared to methods in Ref. [6–8]. However, careful resource estimates in this direction led to costs which are still a few orders of magnitude larger than what is required for Ref. [6–8]. Therefore, in order to use these methods, one would likely need a larger fault-tolerant quantum computer.

The purpose of this paper is to explore an alternative approach to study properties of materials on a quantum computer. The approach we investigate will be variational in nature [19], employing the deep entanglement renormalization ansatz (DMERA) [20], a variant of Vidal's multi-scale entanglement renormalization ansatz [21]. While there may be other ansatzes that one may choose for this purpose, we investigated DMERA primarily for two reasons. First of all, it is well-known that, for the purpose of measuring expectation values of local observables, the DMERA circuit can be compressed to a circuit applied to a constant number of qubits, independent of the size of the system [20–22]. The accuracy of the computation increases progressively as the number of qubits being used increases. Therefore, even with a small quantum computer, one can hope to approximate ground state properties of a much larger system. Secondly, the DMERA ansatz is known to be suitable to describe a large class of phases of matter, such as topologically ordered systems [23, 24] and systems near the point of quantum phase transition [21, 25, 26]. These studies suggest that DMERA is an ansatz capable of approximating a wide class of physical systems.

The main advantage of approaches like DMERA is that it leverages the existing progress made in the tensor network literature [21, 27, 28]; see Ref. [29] for a review.

These are classical computational methods whose underlying complexity is quantified in terms of the "bond dimension," related to the number of parameters in the ansatz. Often times a better approximation to the ground state properties can be obtained by increasing the bond dimension. It was realized in Ref. [20, 30, 31] that the tensor network approach can be employed on a quantum computer, used as an ansatz for the variational quantum eigensolver (VQE) [19]. The key observation was that the number of qubits and gates needed on the quantum computer needed to compute physical properties of interest, e.g., correlation functions and energy, scales *logarithmically* with the bond dimension. The practical bottleneck in tensor network approaches has been an unfavorable scaling in the bond dimension. Since a quantum computer can improve this scaling substantially, the tensor network approach can be potentially sped up substantially by a quantum computer.

In the case of DMERA, the speedup one can hope to obtain can be quantified in the following way. The "size" of the DMERA ansatz can be quantified in terms of two parameters, D and n, which are related to the total depth of the circuit and the width of the circuit as Dn and 2^{dn}, respectively. (Here d is the number of spatial dimensions.) A quantum computer can compute expectation values of local observables in $\mathcal{O}(Dn)$ time using $\mathcal{O}(D^d)$ qubits. In contrast, a classical computer requires memory and computation time that scale *exponentially* with D^d. The dependence on n is the same in both cases, but the exponential dependence on D^d makes a classical simulation challenging. Therefore, one can expect a substantial speedup in estimating the expectation values of local observables.

In spite of these advantages, there are obvious concerns one can raise for this approach. Existing studies in variational quantum algorithms have shown that there are several potential issues one must be wary about. The first issue is that often the sheer number of measurements needed to estimate the energy can be quite large in practice [32]. Another important issue is that the gradient tends to vanish exponentially with the circuit depth [33]. Understanding whether these issues persist in DMERA shall be the primary goals of this paper.

To that end, we carry out a detailed study on the prospect of using DMERA for interacting fermions in two spatial dimensions, focusing on the cost needed for a fault-tolerant quantum computer. We do so by first critically analyzing the issue of barren plateaus [33]. "Barren plateau" refers to a phenomenon in which the gradient of a parameterized quantum circuit decays exponentially with the circuit depth. We compute the typical value of the gradient and observe that its magnitude decays much slower than what was suggested in structureless ansatzes. Specifically, we find that the variance depends greatly on the location of the gates. For instance, the gates that appear towards the end of the circuit has a variance decaying exponentially *only* in D, but not in n. The gates that appear early, on the other hand, does decay exponentially in both D and n. This is similar to the conclusion recently made in Ref. [34]. Though the assumptions we make on the gates are different, our conclusions agree.

Secondly, we perform a detailed resource estimate for estimating a variational energy up to a fixed precision. We obtain a bound sharper than the one derived in Ref. [31] and calculate the exact numbers numerically. The resource estimate was performed for two different schemes for measuring the energy, which are complementary to each other. We refer to the first approach as *incoherent sampling*, which is based on a sampling of the expectation values of the local terms in the Hamiltonian. An encouraging fact about this approach is that the number of qubits needed is substantially lower than the approaches based on QPE.

However, because this approach has a quadratic scaling in the inverse precision, the number of gates needed is greater compared to QPE-based approaches. We have estimated these numbers in detail for a concrete physical system: the Fermi-Hubbard model in two spatial dimensions. We chose a relatively large system size of 64×64, as a proxy for understanding the performance of this method applied to the case of studying $BiFeO_3$ film. Comparing the gate count estimate for the same model using the state-of-the-art approach based on QPE [35, 36], we find the incoherent approach to require a larger number of T gates. Specifically, we find the number of T-gates needed is likely at least an order of magnitude larger than that required for QPE. Therefore, while the incoherent approach may be more advantageous for a small fault-tolerant quantum computer, for a large enough quantum computer it is expected to be inferior compared to the conventional QPE-based approaches.

Fortunately, Heisenberg scaling can be restored under a reasonable assumption on the choice of gates used in the DMERA ansatz. We show that, if the gates chosen in the DMERA circuit is periodic with some finite period, one can exploit this to improve the complexity of estimating the variational ground state energy. Specifically, we improve the dependence on inverse precision from $\mathcal{O}(1/\epsilon^2)$ to $\mathcal{O}(1/\epsilon)$ whilst only increasing the qubit count from $\mathcal{O}(1)$ to $\mathcal{O}(\log N)$, where N is the system size. Since such periodicity is present in many practical calculations [22, 26], we expect this assumption to be a realistic one. This is a novel amalgamation of VQE [19] and amplitude estimation [37], which may be useful in a broader context.

While carrying out these resource estimates, we obtain the following additional results, which may be of independent interest. We introduce a novel compilation technique that can reduce the number of qubits substantially. We observe up to two-fold reduction in the number of qubits for DMERA [Fig. 11]. While our main motivation is to reduce the qubit count for DMERA, the algorithm itself is applicable to any quantum circuits. Moreover, we introduce novel compilation techniques for a fermionic variant of DMERA. This is an ansatz defined in terms of a fermionic gates. A similar ansatz was

already introduced and studied in Ref. [38]. In particular, it is well-known that the efficient contractibility of MERA remains intact for this extension. The fact that this efficient contraction algorithm can be ported to a quantum circuit (with a substantially lower complexity) can be guaranteed on general grounds [39]. However, we provide a more explicit and optimized algorithm which achieves the same goal. Our novel contribution is an explicit realization of a physical operation (in terms of one- and two-qubit gates) that "resets" the fermion occupation number, using the Jordan-Wigner transformation [40]. Using this technique, we were able to extend our resource estimate for qubits to fermions.

The rest of the paper is structured as follows. In Section II, we review DMERA [20, 21]. In Section III we review the concept of past causal cone and elucidate its relationships with the parameters that quantify the size of a DMERA ansatz. In Section IV, we estimate the typical value of the gradient in DMERA, quantifying it in terms of the parameters that determine the size of DMERA. In particular, we show that the gradient decays much slower than what one would naively expect, sidestepping the barren plateau problem [33]. In Section V, we perform a resource estimate for the number of gates and qubits. In Section VI, we describe a quantum algorithm that improves the asymptotic cost of the algorithm in Section V in terms of gate count, with a logarithmic increase in the qubit count. We conclude with a set of open problems in Section VII.

II. DMERA

Let us begin by introducing the central piece of our work: *deep multi-scale entanglement enormalization ansatz*, or DMERA for short [31]. DMERA is a variant of the well-known multi-scale entanglement renormalization ansatz (MERA) [21], which is a variational ansatz that has been used successfully to study ground state properties of quantum many-body systems at and away from quantum critical points [21, 41]. The key advantage of MERA, compared to other variational ansatzes, is that it can produce a power-law decay of correlation, a feature that is commonly present in quantum many-body systems at their quantum critical points.

If we can study the quantum critical points of a general locally interacting Hamiltonian, we would be able to obtain a wealth of information. Quantum critical points often contain a wide variety of universal data that characterizes the underlying phase, such as the critical exponent and the central charge of the underlying theory. Such data can often pinpoint the underlying theory to a select few if not one, which can then constrain the phases that lie on both sides of the transition in a highly nontrivial way [42]. As such, a successful application of MERA to higher dimensions can potentially lead to tremendous breakthroughs in understanding exotic phenomena that can occur in interacting quantum many-body systems, such as high-temperature superconductivity [43].

Unfortunately, while MERA has been used successfully in one spatial dimension (1D) [21, 41], its usage in two spatial dimensions (2D) [41, 44, 45] and above has been relatively limited. This is mainly due to the fact that the computational cost of using MERA can be prohibitive in practice. Specifically, suppose we want to study a system consisting of N particles and we are interested in estimating the computational cost of measuring an expectation value of a local observable. This cost scales merely *logarithmically* with N [41], but the "proportionality constant" in front of the logarithm is too enormous in practice.

To understand where this large constant comes from, it is helpful to consider a simple toy example of MERA, which we describe in Figs. 1. As one can see in these figures, MERA is simply a quantum circuit applied to a fixed product state such as $|0\dots0\rangle$. Each unitary in Fig. 1 generally acts on a $(\chi \times \chi)$-dimensional Hilbert space, where $\chi = 2$ and 4 in Fig. 1(a) and (b) respectively. Assuming that the unitaries in these circuit diagrams can take any value in $U(\chi^2)$, the ansatz in Fig. 1(a) forms a strict subfamily of Fig. 1(b). Therefore, a smaller bond dimension equates to studying a smaller family of quantum states.

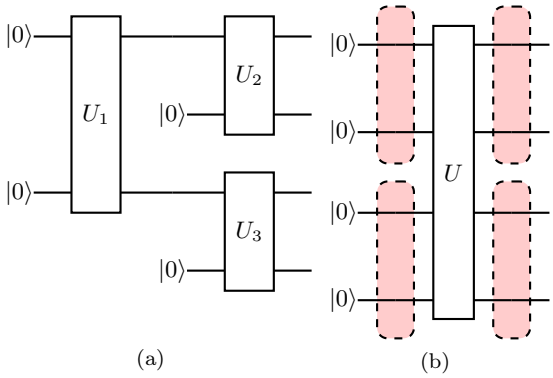


Figure 1. (a) 1D MERA with bond dimension $\chi = 2$ over 4 qubits. (b) An arbitrary unitary acting on 4 qubits. Formally, this can be viewed as a 1D MERA with bond dimension $\chi = 4$ over 4 qubits by viewing U as a unitary acting on two copies of $\chi^2 = 4$-dimensional Hilbert spaces (red).

This parameter χ – also known as the *bond dimension* – is precisely the one that determines the proportionality constant in front of $\log N$. For 1D MERA, using the state-of-the-art method, the dependence can grow as $\mathcal{O}(\chi^5)$ [46]. In 2D MERA, depending on the geometry of the quantum circuit, one gets anywhere between $\mathcal{O}(\chi^{16})$ to $\mathcal{O}(\chi^{24})$ [41, 44, 45]. While the computational cost of MERA is "polynomial" in the bond dimension, the degree of the polynomial is so large that it is very difficult to make it practical. To obtain a more accurate approximation of ground states, χ must be increased. However, this enormous degree is what prevents us from using a

large value of χ

Fortunately, a near-term quantum computer can potentially come to the rescue to alleviate this challenging problem [31]. Here, the computational cost merely grows *logarithmically* with χ, and as such, the enormous constant prefactor can be significantly reduced on a quantum computer.

Below, we will provide a systematic approach to constructing DMERA. Roughly speaking, DMERA can be thought of as a variant of MERA in which the unitaries acting on $(\chi \times \chi)$-dimensional Hilbert space are broken down into a sequence of two-qubit gates acting on $\lceil \log_2(\chi) \rceil$ qubits.

More formally, we can define DMERA as follows. Instead of explicitly drawing the circuit, we will define DMERA in terms of two abstract procedures: *fine-graining* and *finite-depth quantum circuit*. Mathematically, fine-graining refers to a process in which a general quantum state $|\psi\rangle$ is embedded into a larger Hilbert space by appending a simple state such as $|0\dots0\rangle$. Finite-depth quantum circuit means a unitary consisting of geometrically local gates with depth that is independent of N. DMERA is then defined as an alternating sequence of finite-depth quantum circuit and fine-graining [Fig. 2].

Note that there is a great deal of freedom in choosing the fine-graining procedure. The number of newly introduced qubits, as well as the way in which the qubits are arranged with respect to another, are the important choices that can affect the performance of DMERA. Likewise, there are different choices one can make about finite-depth quantum circuits, such as the underlying gate set and where the gates act within each layer of the circuits.

To be quantitative in our statements, we will have to make some choices, which we describe as follows[47]. We shall focus on physical systems in which the qubits are arranged on a (hyper-)cubic lattice of dimension d. We will only consider $(r_1, \dots, r_d)$-*regular* fine-graining procedures, with $r_i = 2$, meaning that the qubits appearing prior to the fine-graining procedure are embedded evenly in a larger lattice elongated in the i'th direction by a factor of $r_i = 2$ for $i = 1, \dots, d$. We shall refer to these constants as *expansion coefficients*. As for the finite-depth unitaries, we shall decompose them into unit-depth quantum circuits, each of which consists of non-overlapping two-qubit gates acting on nearest-neighbors in a fixed direction. For instance, if we fix the direction to $\hat{x}$, we can consider a tensor product of unitaries acting on $(2i, \dots)$ and $(2i+1, \dots)$ for $i \in \mathbb{Z}$, where the rest of the coordinates are the same and can take an arbitrary integer value within the hypercube considered. Alternatively, one can consider a tensor product of unitaries acting on $(2i-1, \dots)$ and $(2i, \dots)$.

To summarize, DMERA is defined in terms of a sequence of fine-graining procedures and finite-depth quantum circuits. We can formally write this as

$$|\Psi\rangle = \mathcal{U}_n \mathcal{F}_N \dots \mathcal{U}_1 \mathcal{F}_1, \tag{1}$$

where the $\mathcal{F}_i$ and $\mathcal{U}_i$ are the i'th fine-graining procedure and finite-depth quantum circuit. Here n is a parameter that quantifies the size of the system. Specifically, let L be the linear size of the system. Assuming that the DMERA is r-regular, the linear size is related to n as $L = r^n$. Thus, we will focus on understanding the computational cost of using DMERA in terms of the following three parameters.

n : Number of scales
r : Expansion coefficient
D : Circuit depth of each finite-depth quantum circuit

To be more precise, we will assume that the underlying DMERA is r-regular and moreover assume that every $\mathcal{U}_i$ has a circuit depth of D with a sequence of unit-depth quantum circuits described in the paragraph above.

III. PAST CAUSAL CONE

While one can in principle directly implement the entire set of finite-depth quantum circuits and fine-graining procedures, that will not be the most economical choice. Instead, given an observable $\hat{O}$, one can simply discard all the gates that do not affect the expectation value of $\hat{O}$. What remains is the *past causal cone* of $\hat{O}$ [41, 48] [Fig. 5].

In order to understand how to obtain the past causal cone, it is helpful to think in the Heisenberg picture. Specifically, recall that the DMERA wave function $|\Psi\rangle$ can be prepared by an alternating sequence of fine-graining and finite-depth quantum circuits, denoted as $(\mathcal{F}_i : i = 1, \dots, n)$ and $(\mathcal{U}_i, i = 1, \dots, n)$, respectively:

$$|\Psi\rangle = \mathcal{U}_n \mathcal{F}_n \dots \mathcal{U}_1 \mathcal{F}_1. \tag{2}$$

In particular, $\mathcal{F}_1 = |0\dots0\rangle$ is simply a preparation of a product state over 2^n qubits. In the Heisenberg picture, an observable "evolves" in the following way:

$$\hat{O}_{k+1} = \mathcal{F}_{n-k}^\dagger \mathcal{U}_{n-k}^\dagger \hat{O}_k \mathcal{U}_{n-k} \mathcal{F}_{n-k} k = 0 \mod D. \tag{3}$$

for $k \in \{0, \dots, n-1\}$. Given a support of $\hat{O}_k$, one can bound the support of $\hat{O}_{k+1}$ by removing the gates in $\mathcal{F}_{n-k}$ and $\mathcal{U}_{n-k}$ in the following way. To be more specific, note that $\mathcal{U}_{n-k}$ is a depth-D quantum, consisting of D layers of unit-depth quantum circuit. Within each unit-depth quantum circuit, we can remove the gates whose support is disjoint from the Heisenberg-evolved operator up to that point, since for any such u it commutes with the operator. Reparing this D times, we obtain a support that is expanded from that of $\hat{O}_k$ [Fig. 3(a), (b)]. Next, the fine-graining procedure in the Heisenberg picture becomes a *coarse-graining* procedure, reducing the support of the observable [Fig. 3(c)].

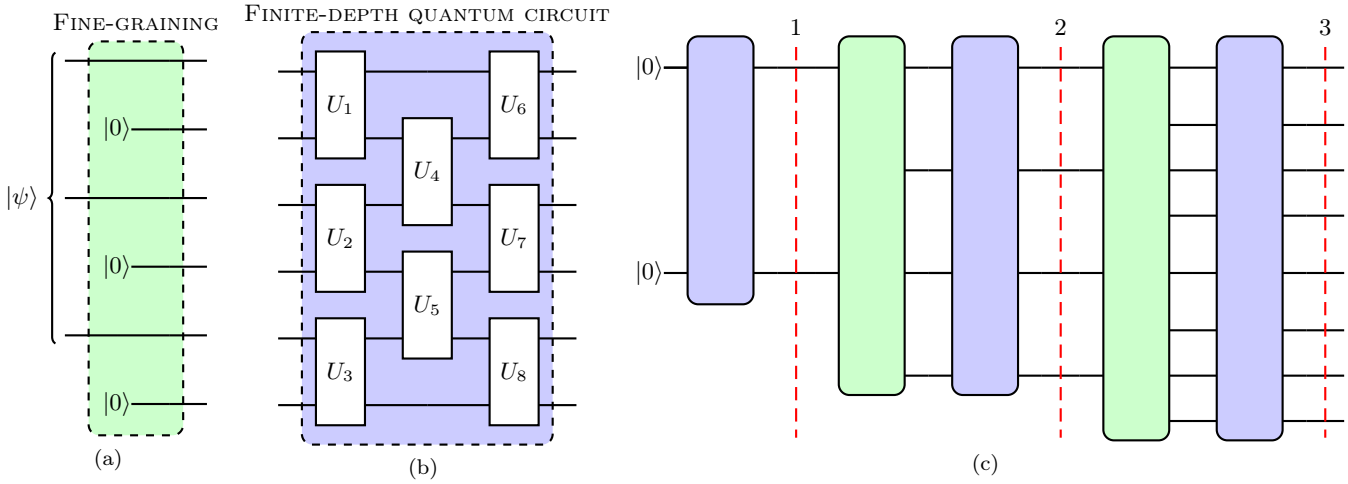


Figure 2. (a) Fine-graining a three-qubit state $\langle\psi|$ to $\langle\psi|\langle 000|$ (b) A depth-3 quantum circuit (c) DMERA as an alternating sequence of finite-depth quantum circuit and fine-graining.

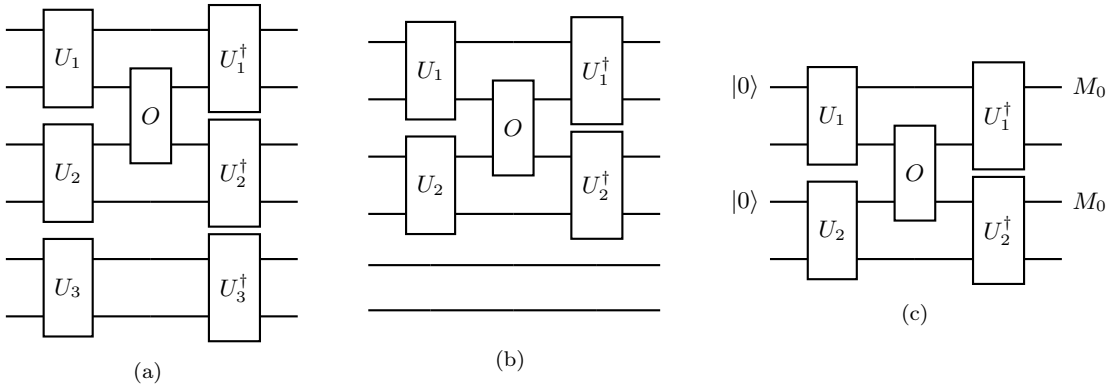


Figure 3. (a) A unit-depth circuit applied to an observable in the Heisenberg picture. (b) Some of the unnecessary gates can be removed from (a), resulting in a reduced circuit. (c) Fine-graining step in the Schrödinger picture becomes the coarse-graining step in the Heisenberg picture. Here M_0 is a post-selection of $|0\rangle$.

Now we can bound the size of the support of $\hat{O}_k$ as follows. For simplicity, let us assume that $\hat{O}_k$ is supported on a ball of radius R_k. Assuming that the depth of $\hat{U}_i$ is bounded by D for all i, the support of $\mathcal{U}_{n-k}^\dagger \hat{O}_k \mathcal{U}_{n-k}$ is bounded by $R_k + D$. Assuming that the fine-graining procedure is $(r, \ldots, r)$-regular, after coarse-graining the radius is bounded by $\sim R_k/r + D/r$. Therefore, we obtain the following recursive inequality:

$$R_{k+1} \lessapprox \frac{R_k}{r} + \frac{D}{r}. \tag{4}$$

As $k \to \infty$, we can expect $R_{k+1} \approx R_k$, yielding

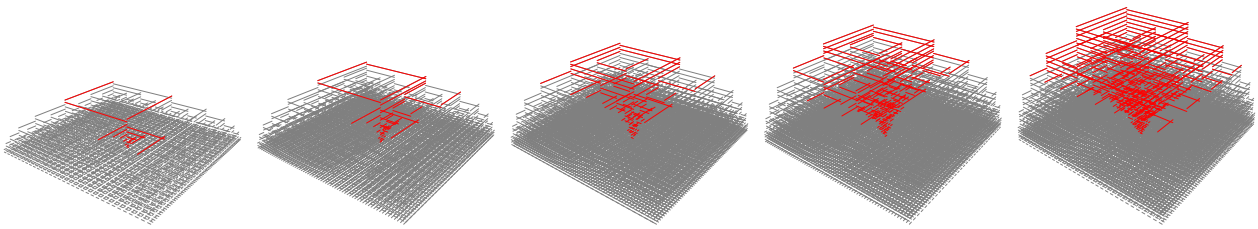
$$R_k \approx \frac{D}{r-1}. \tag{5}$$

Upon evolving the operator in the Heisenberg picture, the radius is again expanded at most by an additive factor of D, yielding the radius of the past causal cone of $Dr/(r-1)$. Thus, the size of the support of $\hat{O}_k$ remains as $((Dr/(r-1))^d$ for all k. The gates that map $\hat{O}_k$ to $\hat{O}_{k+1}$ are also all supported on a ball of radius $\mathcal{O}(Dr/(r-1))$.

By converting this reduced circuit back into the Schrödinger picture, we finally obtain the past causal cone of the observable $\hat{O}$ [Fig. 5(b)]. Moreover, by noting that some of the qubits are no longer in use, one can reduce the width of the circuit to be independent of n [Fig. 5(c)]. Thus, the circuit depth and the width of the past causal cone of DMERA for every local observable is

$$\boxed{\begin{aligned}&\text{Depth} = \mathcal{O}(nD),\\ &\text{Width} = \mathcal{O}((Dr/(r-1))^d).\end{aligned}} \tag{6}$$

This entire process amounts to converting $\mathcal{U}_k$ and $\mathcal{F}_k$ to the versions in which the gates whose supports are disjoint from the support of the Heisenberg-evolved operator. We shall explain this in more detail in Section III A.



(a) $D = 2$ (b) $D = 3$ (c) $D = 4$ (d) $D = 5$ (e) $D = 6$

Figure 4. Examples of past causal cone for 2D DMERA. The black lines represent two-qubit gates appearing in the DMERA and the red lines represent the past causal cone of an observable at site $(x, y) = (16, 16)$. Here n is chosen to be 6.

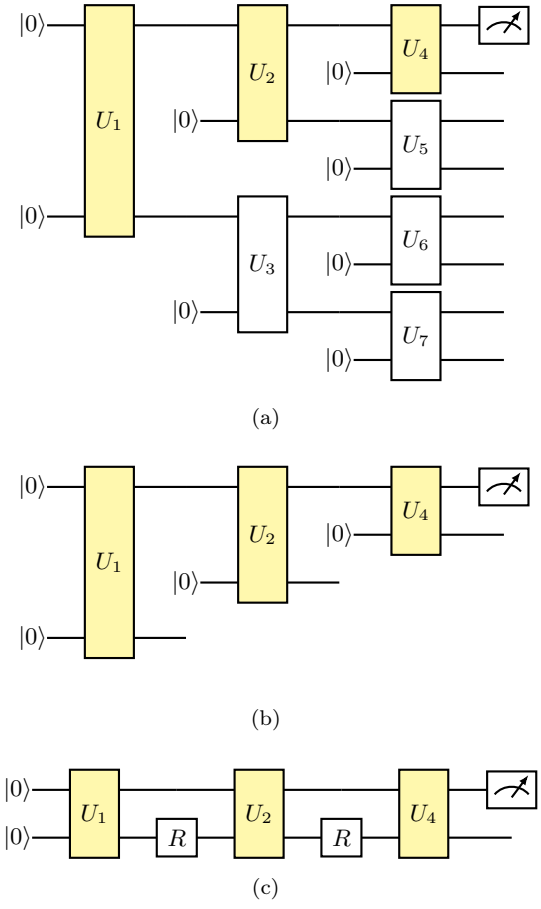


Figure 5. (a) An example of a 1D DMERA with $n = 3$ and $D = 1$. Note that only the yellow gates can affect the measurement outcome of the top-most qubit. (b) After removing the irrelevant gates and qubits, we obtain the past causal cone. (c) By noting that some of the qubits are no longer in use, one can simply feed them into $|0\rangle$, leading to a width-reduced version of the past causal cone. Here R resets the qubit to the $|0\rangle$ state.

A. Transfer matrix

The past causal cone discussed in Section III can be more conveniently described in terms of *transfer matrices*. Recall that, in constructing the past causal cone, we remove the gates whose supports are disjoint from the support of the Heisenberg-evolved operator. Given an operator $\hat{O}$, this entails changing the $\mathcal{F}_i$ and $\mathcal{U}_i$ to the ones in which these gates are removed, which we shall denote as $\mathcal{F}'_i$ and $\mathcal{U}'_i$ below. Then the expectation value of $\hat{O}$ can be written as

$$\Phi_n \circ \ldots \circ \Phi_1(\hat{O}), \tag{7}$$

where Φ_k can be obtained from the composition of $\mathcal{U}'_{n-k}$ and $\mathcal{F}'_{n-k}$, specifically

$$\Phi_k(\cdot) = {\mathcal{F}'_{n-k}}^{\dagger} {\mathcal{U}'_{n-k}}^{\dagger}(\cdot)\mathcal{U}'_{n-k}\mathcal{F}'_{n-k}. \tag{8}$$

We refer to $\{\Phi_k\}$ as the set of transfer matrices.

In principle, all the transfer matrices can be different, in particular, if all the gates in the DMERA are chosen differently. However, often times it is helpful to choose the gates in a translationally invariant way. Such a choice is motivated from existing numerical studies and known analytic constructions of MERA [25, 26]. If that is the case, a more regular pattern arises within the transfer matrices.

To explain how this pattern works, it will be helpful to provide an example of the one-dimensional DMERA. We have computed the past causal cone of nearest-neighbor observables, which are then shifted by an integer multiple of the unit spacing [Fig. 6]. As one can see, all the past causal cones within each scale can be obtained from a shift and a restriction of a transfer matrix on a sufficiently large interval. The same conclusion applies to 2D DMERA.

The shape and the size of the past causal cone play a fundamental role in the resource estimate for DMERA. For each scale s, we denote the maximum width as w_s. The maximum of w_s over s is the number of qubits we need to run DMERA.

IV. GRADIENT SCALING

In this Section, we study the typical value of the gradient when the gates chosen in the DMERA are random. To model whether the ansatz generates a barren plateau,

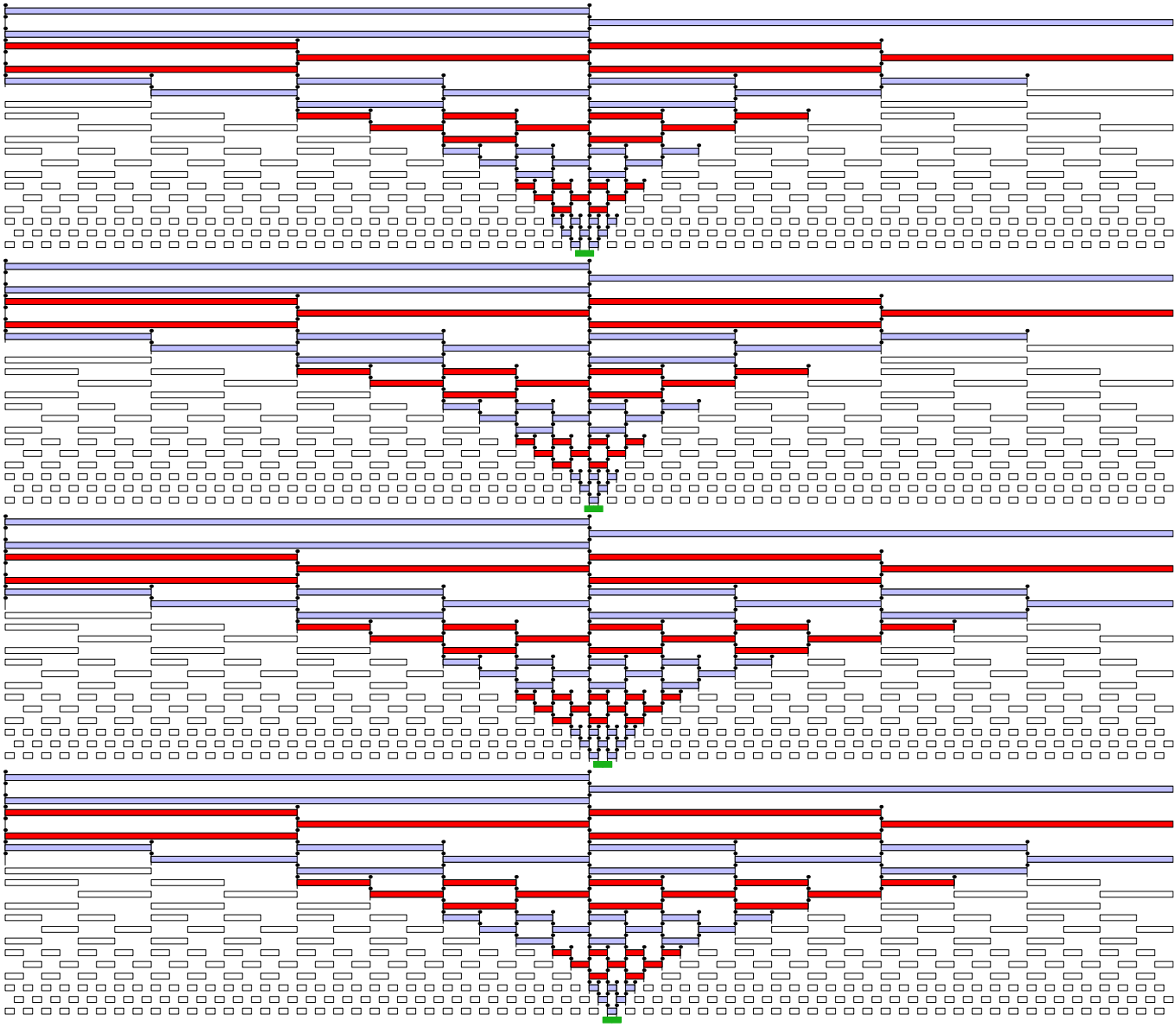


Figure 6. Examples of past causal cone for nearest-neighbor observables on $(n = 8, D = 3)$-DMERA. The nearest-neighbor observable is represented as a green rectangle. Different scales are colored in red/blue depending on whether the scale is odd/even; this is to clearly delineate the different scales. One can see that the transfer matrix within each scale can be obtained by shifting a sufficiently large transfer matrix within that scale and then further restricting that transfer matrix to a smaller set of qubits.

namely a gradient that scales polynomially in some variable only for an exponentially small volume of the space of all circuits, we describe our circuit as a product of Haar random two-qubit unitaries,

$$\prod_{i=N}^{1} \mathcal{U}_i \left|\vec{0}\right\rangle$$

and take

$$E = \langle H \rangle = \langle 0| \prod_{i=1}^{N} \mathcal{U}_i^\dagger) H (\prod_{i=N}^{1} \mathcal{U}_i) |0\rangle$$

. Each unitary $\mathcal{U}_i$ is composed itself of a set of unitaries $\mathcal{U}_i = \prod_j U_j$ where j runs over the total number of gates at level i. For simplicitly, we will simply compress all indices and represent our circuit as a long list of unitaries $\{U_i\}$ and denote the number of such unitaries as N for the purposes of this section. We are interested in taking gradients of this energy with respect to the individual unitaries U_i. We can define a gradient of one of the gates U_k by perturbing it with a Pauli operator P, ie replacing

U_k with $U_k \exp\{-i\epsilon P\}$, and computing

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}(\langle 0|(\prod_{i=1}^{k-1}U_i^\dagger)e^{-i\epsilon P}(\prod_{i=k}^{N}U_i^\dagger)H(\prod_{i=N}^{k}U_i)e^{i\epsilon P}(\prod_{i=k-1}^{1}U_i)|0\rangle - \langle 0|(\prod_{i=1}^{N}U_i^\dagger)H(\prod_{i=N}^{1}U_i)|0\rangle)$$

For convenience, we define $|0'\rangle = \prod_{i=k-1}^{1}|0\rangle$ and $P_j' = \prod_{i=k}^{N}U_i^\dagger P_j \prod_{i=N}^{k}U_i$, and take the Hamiltonian to be some sum of Pauli operators $H = \sum_j \alpha_j P_j$. With this convention the above can be rewritten as

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\langle 0'|(e^{-i\epsilon P}\sum_j \alpha_j P_j' e^{i\epsilon P}) - \sum_j \alpha_j P_j'|0'\rangle$$

Expanding the sum over j we get

$$\begin{aligned}
\partial_{P,k}E &= \sum_j \alpha_j \partial_{P,k}\langle P_j\rangle \\
\partial_{P,k}\langle P_j\rangle &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\langle 0'|(|e\rangle^{-i\epsilon P}P_j' e^{i\epsilon P} - P_j')0' \\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\langle 0'|P_j' - i\epsilon(P_j'P - PP_j') + \mathcal{O}(\epsilon^2) - P_j'|0'\rangle \\
&= \mathrm{Im}(\langle 0'|\left[P_j', P\right]|0'\rangle)
\end{aligned}$$

Our interest is the distribution of $\partial_{P,k}E$ over the space of circuits generated by Haar random two-qubit unitaries U_i. We can see immmediately that $\langle \partial_{P,k}E\rangle_{\text{Haar}} = 0$, as for every circuit which generates a positive expectation value for any term one can modify it to produce the negated value. Our question then is to understand the variance of $\partial_{P,k}E$, which since the expectation value is zero becomes just

$$\begin{aligned}
\mathrm{Var}(\partial_{P,k}E) &= \langle \partial_{P,k}E\rangle_{\text{Haar}} = \sum_{i,j}\alpha_i\alpha_j\langle \partial_{P,k}P_i\partial_{P,k}P_j\rangle_{\text{Haar}} \\
&= \sum_{i,j}\alpha_i\alpha_j\left\langle \mathrm{Im}(\langle 0'|[P_i', P]|0'\rangle\,\mathrm{Im}(\langle 0'|\left[P_j', P\right]|0'\rangle\right\rangle_{\text{Haar}}
\end{aligned}$$

Note that the expectation value from Haar-random circuits is equal to the expectation value from random Clifford circuits with identical connectivity, and thus we can say that

$$\mathrm{Var}(\partial_{P,k}E) = \sum_{i,j}\alpha_i\alpha_j\left\langle \mathrm{Im}(\langle 0'|[P_i', P]|0'\rangle\,\mathrm{Im}(\langle 0'|\left[P_j', P\right]|0'\rangle\right\rangle_{\text{Clifford}}$$

We note some key properties of the terms $\partial_{P,k}\langle P_j\rangle = \langle 0'|\left[P_j', P\right]|0'\rangle$. Thanks to shifting to drawing random Clifford circuits, we note that P_j' is now another Pauli string as opposed to some general operator, as Cliffords map Paulis into Paulis with some phase. Specifically, P_j' is some Pauli string that is supported on the past causal cone of P_j at the point in the circuit corresponding to gate U_k. Second, $\left[P_j', P\right]$ thus is either 0 if operator P commutes with P_j', or it is equal to $2P_j'P$. Thus, the value of $\partial_{P,k}\langle P_j\rangle$ for any given circuit is 0 if P and the image of P_j defined over that operator's PCC at gate U_k commute, or it is equal to twice $\langle 0|P_j''|0\rangle$ where P_j'' is the Pauli operator corresponding to $P_j'P$ projected back through to the beginning of the Clifford circuit. This evaluates to 0 if P_j'' has any X or Y Paulis, otherwise it evaluates to 1.

Thus, we can recast the task of computing the variance of the gradient of the energy to simply asking what the probability is that a) P_j' and some P commute and b) that P_j'' has no X or Y Pauli operators, and what the correlation is between that probability for different Hamiltonian terms P_j.

A. Upper bound of the variance of the gradient

We can upperbound the probability by first taking the norm of each term in the variance, to discard consideration of the phase coming from the Clifford circuit. Second, we can pull out a factor of 4 and treat each term in the product as a simple binary variable (corresponding in value to whether or not both $\left[P_j', P\right] \neq 0$ and if P_j'' contains no X or Y Paulis). Third, we note that $\langle x_i x_j\rangle$ for some binary variables x_i is strictly smaller than $\max_{i,j}(\langle x_i\rangle, \langle x_j\rangle)$. Thus, we can upperbound the variance by asking this question: What is the probability that both our conditions $\left[P_j', P\right] \neq 0$ and $\{\sigma_X, \sigma_Y\} \notin P_j''$ are satisfied for a random Clifford circuit? If this probability decays exponentially in either the DMERA depth D or the number of layers n then we are guaranteed to have a barren plateau along that parameter and the circuit will rapidly become untrainable as we scale that parameter.

The last important feature to note is that a random 2-qubit Clifford operation maps $II \to II$ while transforming any other Pauli string to a random Pauli string (that is not II). We can then see that the action of the circuit on a Pauli string operator P is the identity on all gates that do not interact with P, while scrambling the Pauli string otherwise.

Let us examine the case with two gates G_1 and G_2 acting on a set of 3 qubits q_1, q_2, and q_3, where q_1 and q_2 are inside our Pauli string and q_3 is just outside it. Assume that the Pauli string is not II on q_1, q_2, then G_1 will scramble the string and the operator on q_2 will be I with probability 3/15. If this is the case, then the gate G_2 will be acting on II and not expand the Pauli string.

There are two categories of operators, I and $\{X, Y, Z\}$, and the action of random Cliffords renders $\{X, Y, Z\}$ equally likely at every stage, and we can simplify our discussion by only looking at a binary choice between these two classes, call them I and V. Finally, we note that for our purposes of placing a bound on the variance, we can ignore the question of whether the $\left[P_j', P\right] = 0$ or not as it is at most a constant factor (since P is just a two-body operator), and moreover that since the Clif-

ford circuit randomizes P_j' we can drop dependence on the specific Pauli P.

Now, bounding the variance is equivalent to asking what is the distribution of Pauli operators in P_j''.

We can model the evolution of P_j through the random Clifford circuit as a random walk. "In the bulk" of the past causal cone, individual operators are I with a probability of 3/15, as each gate acting in the bulk has a probability of 3/15 of setting each of the operators to I (but not both). Therefore, the probability of all operators in P_j'' being I or Z is driven by the size of the boundary of support for P_j''.

As demonstrated above, a gate acting at the boundary of the PCC can either expand it (with probability 12/15) or contract it (with probability 3/15). Extinction of P_j's PCC is impossible as random Cliffords cannot map any operator to II except II. We can treat the movement of the boundary as a kind of random walk. In a single dimension, take the width of the effective PCC after each round of nearest neighbor interactions in the DMERA as a random walker that can move $-2, -1, 0, 1$, or 2 steps on each move, bounded from below at 2 (as P_j is a two-body operator) with probabilities of $1/25, 2/25, 6/25, 6/25$, and $9/25$ respectively. To see this, simply count the possibilities and probability of each outcome from the fact that each edge has a 3/15 chance of placing an I on the boundary and thus pulling the PCC inward, a 3/15 chance of placing some other operator on the boundary and thus not moving it, or otherwise expanding the PCC.

This random walk will grow with time at an average rate of 4/5 per round, with a standard deviation which grows as the square-root of the number of rounds. After D such rounds, we thus expect the PCC's operator along one dimension to extend to $0.8D$ qubits. After each layer of our DMERA circuit is completed, we then reduce this width by a factor r, and continue to expand.

In expectation, then, the width in a single dimension of P_j'' after pushing P_j backward through an entire random Clifford circuit will be, as one could predict from Eq. 6, $\langle \text{Width} \rangle = 0.8Dr/(r-1)$ with a standard deviation that scales as $\sigma = \mathcal{O}\left(\sqrt{D}\right)$.

Thus, the ball corresponding to the PCC of P_j has a radius of $R \approx 0.4Dr/(r-1)$, and its volume scales as R^d. After each layer of DMERA, an additional volume

$$\mathcal{O}\big((R(1-1/r))^d\big) = \mathcal{O}\big((0.4D)^d\big)$$

with n such levels for a total number of qubits $\mathcal{O}\big((0.4D)^d n\big)$ inside P_j'' in expectation.

However, while on average the volume of the PCC grows linearly in n and thus a large fraction of all circuits will contribute $\mathcal{O}(exp(-n))$ to $\langle P_j \rangle$, at each scale level, label it $s \in [1, n]$ there is the possibility that P_j' will have support only on the $\mathcal{O}\big((0.4D)^d\big)$ nodes which will not be acted upon again later in the circuit – that is, the Pauli operator P_j' will freeze and no longer evolve as we apply the rest of the gates in the PCC of P_j. The probability that this occurs does not scale with n, but rather only with the volume of the PCC after one layer, $\mathcal{O}\big(D^d\big)$. Therefore, there is a constant fraction of all Clifford circuits as a function of n will have a P_j'' with support on these $\mathcal{O}\big(D^d\big)$ nodes.

Moreover, while we expect that the vast majority of circuits will have support on $\mathcal{O}\big(D^d n\big)$ nodes, there is a small probability that P_j' does not percolate through the graph at all – at each layer it has a some chance of not expanding, and thus it has a probability $p = \mathcal{O}(exp(-Dds))$ of simply not expanding through scale s, and remaining in a constant size region of support approximately equal to that of P_j. These terms individually make the largest contributions, as they have a large ($\approx$ constant) probability of containing no X or Y Paulis and that probability decays as 3^{-K} for a Pauli with K non-identity operators (the more general case for the weight of the Pauli string up to scale s would have a probability of $3^{-\mathcal{O}(D^d s)} \ll exp(-\mathcal{O}(Dds))$ for $d > 1$). The expectation value for $\langle 0|P_j''|0\rangle$ is thus dominated by these components. The contributions from these paths can be upper-bounded by assuming the resulting Pauli strings evolve, as in the previous paragraph, to only have support on qubits which are not acted on at longer length scales. However since these P_j''s are small of constant size, this probability is p^{-K} with p on the order of $\frac{r^d-1}{r^d}$ for a DMERA circuit that grows by a factor of r at each scale.

Taking the derivative, then, of a parameter present in scale s, we find the derivative is actually dominated by the cases where P_j' has not expanded in support significantly beyond the size of P_j, and where it does not expand significantly as it evolves to P_j'', ie terms whose contributions decay exponentially with Dds (the number of gates each qubit participates in in a layer) rather than the naive $D^d n$.

Each of these non-I operators is a random Pauli equally distributed over $\{X, Y, Z\}$, then

$$\text{Var}\partial_{k,P} \langle P_j \rangle \leq \mathcal{O}(\exp(-Dds)) \tag{9}$$

and therefore by Chebyshev's inequality the volume of all circuits with polynomially scaling variance for DMERA is exponentially small in the depth D.

Thus, we find that our upper bound allows gradients to scale polynomially with system size, but they decay exponentially with depth parameter D, indicating a barren plateau in D.

B. Lower bound

We can go further than our upper bound by investigating the behavior of $\big\langle \text{Im}(\langle 0'|[P_i', P]|0'\rangle \,\text{Im}(\langle 0'|\big[P_j', P\big]|0'\rangle \big\rangle_{\text{Clifford}}$.

First, we will note that a random Clifford gate acting on a Pauli string will scramble the phase coefficient in front of the Pauli string so long as it does not act trivially (ie as long as the Pauli is not the identity on the support of the gate). The specific properties of the

cross-correlation terms, $i \neq j$, will depend on the parameters α. Moreover, for any given Clifford circuit, one can replace the next gate to act on the support of $P_i' P$ so as to have the opposite phase, resulting in a cancellation of these terms as we average over the ensemble of all Clifford circuits. However, the terms of the form $\alpha_i^2 \left\langle \mathrm{Im}(\langle 0'|[P_i', P]|0'\rangle)^2 \right\rangle_{\mathrm{Clifford}}$ are independent of the sign of the phase, and so as above in the upper-bound case reduces to simply taking an expectation value over an ensemble of terms that are either 0 or 1 over all Clifford circuits, with no cancellation possible as in the $i \neq j$ case.

As in the upper bound case, the scrambling properties of Cliffords acting on Pauli strings allows us to drop specific dependence on P, and we again see the uniformly random transformation over all gates acting non-trivially on a Pauli string allows us to simplify our construction into two classes of operators: at any given location the P_i' has either the identity at any given site or a uniformly random selection over $\{X, Y, Z\}$. One can see immediately that the arguments above, mapping the expansion of the evolution of the support of the P_i' operator through the PCC of P_i acts a random walk, go through unchanged.

We can conclude that for the purposes of a lower bound, we can ignore the contributions of large-support P_i'' circuits (since each circuit can only contribute a positive value to the expectation) and observe that we have a probability that scales as $\mathcal{O}(\exp(-cDds)) = \mathcal{O}(\exp\{-cDd \log N_s\}) = \mathcal{O}\big(\exp\{-cD\} N_s^d\big)$ for a positive value $0 < c < 1$ for the support of the PCC to not expand beyond the region of P_i up to scale s and volume up to scale s labeled N_s.

We see that the upper bound and lower bound both have the same form with respect to D and s, that is, depth and the scale at the location of the target parameter (at most log of the system linear size).

We note that this result may appear to differ with [34], however we note that that reference is concerned with MERA circuits, rather than DMERA. As a result, at each scale, there is only a constant number of operations with neighbors, effectively $D = 1$, and our result here agrees that there should be no barren plateau in that case in system size N. DMERA is a more general case, and we observe an exponential suppression of the variance of the gradients with depth parameter D.

C. Numerical simulations

To demonstrate this, we simulate the evolution of a Pauli operator backward through the DMERA circuit in 2D in a simplified setting, modeling only the effect of random Clifford gates according to the Monte Carlo update rule above, namely that a random Clifford gate will map $II \to II$ with $p = 1$ and fully scramble all other Pauli operators with equal probability. Thus we need only model whether the operator is I or a random Pauli $V \in \{X, Y, Z\}$. We thus can select a pair of qubits denoting a measurement at the end of a DMERA circuit, assign the support of the measurement the state of being some non-I Pauli V (denoting a random X, Y or Z Pauli), and then evolve that state back through a random Clifford circuit to find the effective template for a realization of the Pauli PCC of the measurement, denoted above by P_j''.

Describing these templates for P_j'' in a binary string $\vec{K_{P_j}}$, with 0 denoting an identity gate and 1 denoting a random Pauli, we can then simply evaluate the Hamming weight of the binary string, $\left|\vec{K_{P_j}}\right|$, and look at the distribution of $\left|\vec{K_{P_j}}\right|$ as a function of n and D. We suppress the dependence on P_j for the time being for easy of notation.

A random instantiation of a template Pauli string $\vec{K}$, is found by building a Pauli string with I everywhere $K_i = 0$ and a random value X, Y, or Z everywhere $K_i = 1$. The probability that $\langle 0|\vec{K}|0\rangle = 1$ (it can only equal 1 or 0) is simply $3^{-|\vec{K}|}$. Thus, given an ensemble of random such templates one can estimate the value of $\mathbb{E}[\langle 0|P_j''|0\rangle]_{\mathrm{Clifford}}$, our upper bound on the gradient, as

$$\mathbb{E}[\langle 0|P_j''|0\rangle]_{\text{ Clifford}} = \sum_k Pr(\left|\vec{K_{P_j}}\right| = k) 3^{-k} \tag{10}$$

By sampling many templates for a given measurement operator's support we can estimate our upper-bound the $\partial_{P,k} \langle P_j \rangle$. We can also give an average value of our upper bound over the measurements for each operator.

We simulate this process to construct 10^6 templates for P_j'' for each of 100 randomly selected (with replacement) measurement operators (defined by their supports) for each combination of $n \in \{3, 4, 5, 6, 7, 8\}$ and $D \in \{1, 2, 3, 4, 5, 6\}$ in two dimensions and present the results below.

First, we simply will plot $\left|\vec{K}\right|$ as a function of D and n in Fig. 7, to verify our expectation that on average it scales quadratically and linearly, and it does as expected.

Next, we take the expectation value of our upper bound, as defined in Eq. 10 over all selected measurement operators along with one sigma error bars for said expectation value (determined via a Bayesian bootstrap [49] over the sampled measurement operators) and plot it as a function of n and D in 8, showing a clear exponential decline with D and no n dependence, as theory predicts due to the heavy suppression of high-weight random Pauli strings due to low probability of such strings containing no X or Y Pauli operators. This leaves the region in which our Pauli string does not expand significantly over time over the circuit as the dominant contributor to our upper bound.

We may also break out the individual values for our upper bound for each of the 100 measurement operators for different depths D, for a fixed $n = 8$, though as is

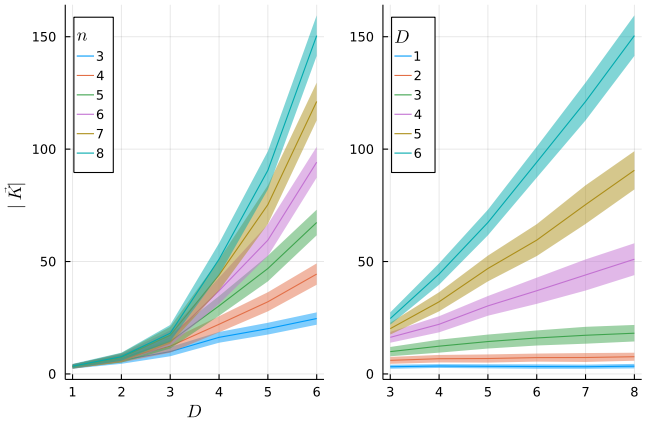


Figure 7. We plot the expectation value and standard deviation of $\left|\vec{K}\right|$ as a function of D and n, and we see our effective past causal cones on average scale in volume polynomially with D but linearly with n, as expected.

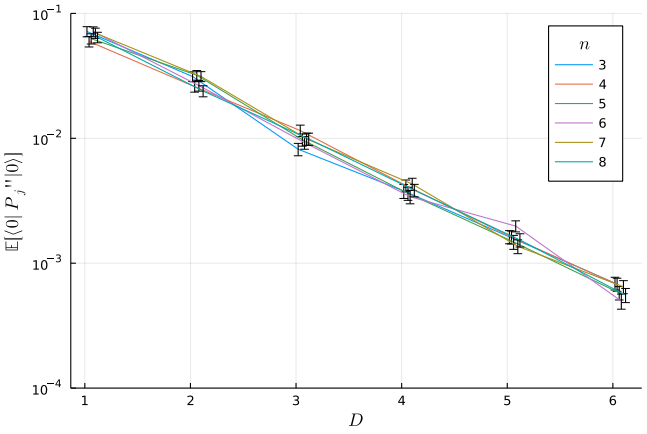


Figure 8. The expectation value of our upper bound on $\partial_{P,k}\langle P_j\rangle$ over 100 randomly selected P_j, ribbon denotes the 1σ of the expectation values, resulting from a Bayesian bootstrap. We see very clearly that there is no relationship between this and n, as predicted, while there is a strong exponential dependence on D.

apparent from Fig 8, there is no significant dependence on n. We show a scatter plot of all such values, sorted by value of the upper bound, in Fig 9. The discrete bands we see with limited variation are attributable to different probabilities of observing the smallest possible $\left|\vec{K}\right|$.

We observe clearly that while there is a large separation across measurement operators, as a function of D there is a consistent exponential decline in the upper bound on $\partial_{P,k}\langle P_j\rangle$. Thus, as predicted, there exists a barren plateau for D, but negligible dependence of the upper bound for the gradient on n.

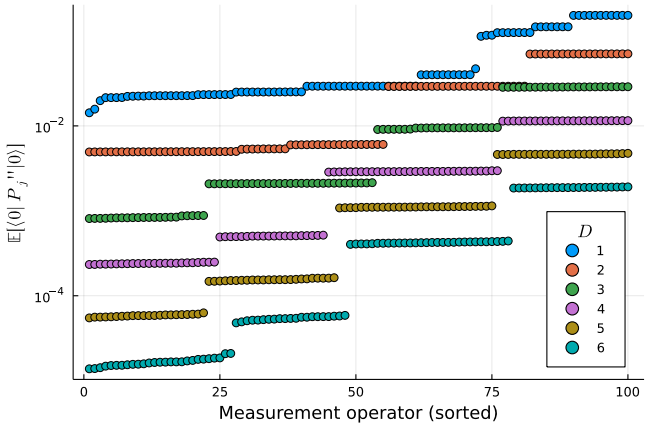


Figure 9. The expectation value of our upper bound on $\partial_{P,k}\langle P_j\rangle$ over 100 randomly selected P_j for various D and $n = 8$, sorted along the x-axis from least to greatest. We see discretized behavior, which we attribute to varying probability to observe the very smallest $\left|\vec{K}\right|$

V. RESOURCE ESTIMATE: INCOHERENT APPROACH

In this Section, we quantify the resource required to run a DMERA circuit in terms of the number of qubits, circuit depth, and the number of one- and two-qubit gates. We assume the standard model of quantum computation in which two-qubit gate between any pair of qubits is possible, and we assume the ability to perform reset of the qubits.

Let us first note that the circuit depth of the (n, D)-DMERA is exactly nD and the number of qubits needed is upper bounded by $\mathcal{O}(D^d)$ and the number of gate is $\mathcal{O}(nD^{d+1})$, where d is the number of spatial dimensions. However, in resource estimation the precise constants matter. We perform a detailed analysis by explicitly computing the past causal cone and estimating its width. Depending on how this circuit is compiled, the depth and the width of the circuit can change. However, the total number of gates remains invariant. We have plotted this number in terms of D for D ranging from 2 to 16 in Fig. 10. For concreteness, we chose the system size to be 64×64, motivated from the physical examples discussed in the introduction.

We have also employed a novel circuit optimization technique that can reduce the number of qubits substantially. We introduce a novel compilation technique which we describe in detail in Appendix A. This technique reduces the overall number of qubits needed with the expense of increasing the depth of the circuit. The comparison between these two are made in Fig. 11.

In a fault-tolerant quantum computer, the relevent figure of merit is the number of non-Clifford gates, such as the T-gate. The precise number can be estimated conveniently in terms of two parameters: (i) the number of two-qubit gates in the past causal cone and (ii) the

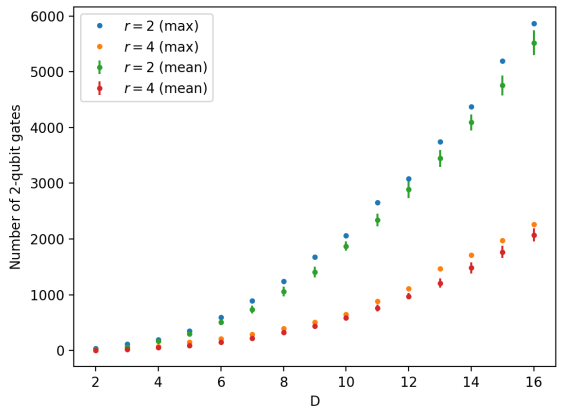


Figure 10. The number of 2-qubit gates used in DMERA in terms of D.

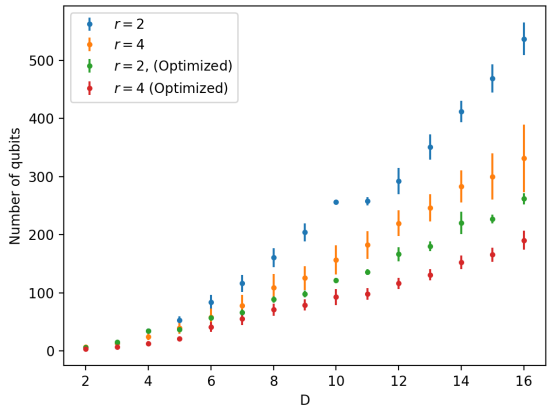


Figure 11. The number of qubits needed for computing past causal cone. Each data point was obtained from 100 randomly generated nearest-neighbor local observables on a two-dimensional lattice. Generally, larger value of r (see Section III) reduces the qubit cost. The compilation technique in Appendix A also led to a noticable reduction in qubit cost.

amortized T-gate cost for each gate. While (i) depends only on D, as shown already in Fig. 10, the second number depends on the choice of the gates. On one hand, if one were to use arbitrary two-qubit gates that define the DMERA ansatz, one would need to employ a decomposition of a general unitary in $SU(4)$ into CNOT gates and single-qubit rotations [50]. This entails using at most 4 arbitrary single-qubit rotations plus 3 single-axis rotations, accompanied with Clifford gates. Since arbitrary single-qubit rotation can be decomposed further into 3 single-axis rotation, using at most 15 single-axis single-qubit rotations suffice. Each such rotation ought to be decomposed further into Clifford and T-gates, with the T-gate cost ranging anywhere between 10 to 50 using standard methods [51–55]. Thus, a pessimistic estimate on this number can be as large as 750 T-gates per individual two-qubit gate.

However, in practice, it is highly unlikely that one would employ such structureless gates. Recent studies showed that simple choice of gates can be often used in practice [56, 57]. In particular, a two-qubit gate chosen from a trotterized form of the local terms of the Hamiltonian is particularly promising [58, 59]. Moreover, often the gates used in these systems are completely arbitrary but have some uniform structure, such as translational invariance [25, 26]. Under those circumstances, it is possible to apply identical rotations over n qubits with the T-gate cost of $4n + \mathcal{O}(\log n \log(1/\epsilon))$, where ϵ is the target precision [60], which in the large n limit, approaches 4 T-gates per rotation.

To summarize, a single coherent run of the DMERA in two spatial dimensions is $n_{2Q}C_T$, where n_{2Q} is the number of two-qubit gates, scaling cubically with D [Fig. 10] and C_T is the amortized T-gate cost of the individual two-qubit gates, ranging between 4 and approximately 750.

A. Sample complexity

In this Section, we estimate the sample complexity of estimating a variational energy density up to an additive precision ϵ. Without loss of generality, consider the Hamiltonian of the following form.

$$H = \sum_{x,y=1}^{L} \sum_{k} \left(\alpha_k O^{(k)}_{(x,y),(x+1,y)} + \beta_k O^{(k)}_{(x,y),(x,y+1)} \right), \tag{11}$$

where $\alpha_k, \beta_k \in \mathbb{R}$ are numerical coefficients and $O^{(k)}_{(x,y),(x+1,y)}$ and $O^{(k)}_{(x,y),(x+1,y)}$ are hermitian operators of norm at most 1. Note that Eq. (11) is the most general form of translationally invariant Hamiltonian consisting of nearest-neighbor terms.

Using translational invariance, the total energy *per site* can be written as:

$$\frac{\langle H \rangle}{L^2} = \sum_{k} \mathbb{E}_{x,y}[\alpha_k \langle O^{(k)}_{(x,y),(x+1,y)} \rangle + \beta_k \langle O^{(k)}_{(x,y),(x,y+1)} \rangle]. \tag{12}$$

Assuming the individual observables $O^{(k)}_{(x,y),(x+1,y)}$ and $O^{(k)}_{(x,y),(x,y+1)}$ are measured, upon optimizing the number of measurements, we obtain the total number of measurements equal to [32]:

$$M = \frac{\left(\sum_k (|\alpha_k| + |\beta_k|)\right)^2}{\epsilon^2}. \tag{13}$$

B. Case study: 2D Fermi-Hubbard Model

Now we apply the resource estimate obtained in Section V and V A to a concrete physical model of interest:

the Fermi-Hubbard Model on a two-dimensional square lattice. Our goal is to compare the resource estimate for our approach against other approaches based on QPE applied to the same model [35, 36]. More precisely, our goal in this Section is to estimate the number T-gates necessary to estimate the variational energy of H', up to an accuracy of half a per cent of the total system energy [35, 36, 61] (corresponding to roughly $0.0051L^2$).

The Hamiltonian of the Hubbard model is defined as

$$H = t \sum_{\sigma\in\{\uparrow,\downarrow\}} \sum_{\langle i,j\rangle} a^\dagger_{i,\sigma} a_{j,\sigma} + h.c. + u \sum_i n_{i,\uparrow} n_{i,\downarrow}, \quad (14)$$

where $a^\dagger_{i,\sigma}$ and $a_{i,\sigma}$ are fermion creation/annihilation operators and $n_{i,\sigma} = a^\dagger_{i,\sigma} a_{i,\sigma}$. Here the summation over $\langle i,j\rangle$ means the summation over i and j which are nearest neighbors on the square lattice.

As pointed out in Ref. [36], Eq. (14) can be rewritten in the following form:

$$H = H' - \frac{u}{2} N + \frac{u}{4} I, \quad (15)$$

where

$$H' = t \sum_{\sigma\in\{\uparrow,\downarrow\}} \sum_{\langle i,j\rangle} (a^\dagger_{i,\sigma} a_{j,\sigma} + h.c.) + \frac{u}{4} \sum_i z_{i,\uparrow} z_{i,\downarrow}. \quad (16)$$

Here I is the identity operator, $N := \sum_i n_{\uparrow,i} + n_{\downarrow,i}$ is the total electron number, and $z_{i,\sigma} := 2n_{i,\sigma} - I$. The last two terms in Eq. (15) commute with H', and as such, they can be simultaneously diagonalized. Moreover, if the underlying state $|\psi\rangle$ is an eigenstate of N, e.g., $N|\psi\rangle = \eta|\psi\rangle$, the eigenvalue of H can be obtained straightforwardly from the eigenvalue of H' by subtracting $u(\eta/2 - 1/4)$. Therefore, for the purpose of computing the ground state energy of H (given a definite value of η), it suffices to compute the ground state energy of H'.

Upon applying the Jordan-Wigner transformation, the first term in H', for each σ and (i,j), can be mapped to an operator of norm at most t. Therefore, the estimate for the sample complexity [Eq. (13)] becomes

$$M \le \frac{\left(2t + \frac{u}{4}\right)^2}{\epsilon^2} \approx \begin{cases} 3.5 \times 10^5 & \text{for } t = 1, u = 4, \\ 6.2 \times 10^5 & \text{for } t = 1, u = 8. \end{cases} \quad (17)$$

Therefore, for achieving the accuracy of a half a per cent of the total sytem energy, the number of samples needed ranges betweem 3.5×10^5 and 6.2×10^5.

As reviewed in Section V, in order to estimate T-gate cost, one must multiply the numbers in Eq. (17) by $n_{2Q} C_T$, where n_{2Q} is the number of two-qubit gates being used and C_T is the amortized cost of the individual two-qubit gates. Even under an optimistic assumption that C_T is very close to the lower bound of 4, the number of T-gates obtained this way easily exceeds the number of T-gates used in the state-of-the-art methods for QPE [36], which approaches at most $\approx 2\times10^6$. Therefore, the incoherent approach is likely to be strictly inferior compared to the QPE-based approaches, in terms of the number of T-gates.

One potential advantage of the incoherent approach, however, is the small number of qubits needed. For instance, in Fig. 11, the number of qubits needed remains at least an order of magnitude smaller than the system size (64×64). Therefore, there may be a merit in using the incoherent approach for DMERA in a small fault-tolerant quantum computer. However, if a large enough quantum computer is available, QPE should be the preferred method of choice.

VI. AMPLITUDE ESTIMATION FOR DMERA

A fundamental limitation of the approach discussed in Section V lies in its scaling in the precision. The sheer number of samples needed to attain a desired accuracy was comparable to the number of T-gates needed for the state-of-the-art approaches based on QPE [35, 36]. The approach taken in Section V is based on samplings of observables, whose scaling in terms of the statistical error ϵ is $\Omega(1/\epsilon^2)$.

It is a well-known fact that, using the quantum amplitude estimation algorithm [37], one can improve the scaling to $\mathcal{O}(1/\epsilon)$. However, a naive implementation of amplitude amplification ends up with some issues. One straightforward way to employ this algorithm is to measure individual local terms in the Hamiltonian (after decomposing them further into a linear combination of Pauli operators). However, the issue is that there are $\Omega(L^2)$ terms in the Hamiltonian. If we were to measure them altogether, it would incur a cost that scales quadratically with L, a substantial increase compared to the logarithmic scaling in L in Section V. If the state prepared by the DMERA circuit is translationally invariant, one could simply pick one of the sites and measure the local term acting nontrivially on that site. However, in practice it is difficult to ensure the translational invariance of the state. Yet another alternative is to sample uniformly randomly over the sites [Eq. (12)], each time employing the amplitude estimation algorithm for the randomly chosen observable. Unfortunately, in this case, we would again restore the $\Omega(1/\epsilon^2)$ scaling because this is a sampling-based approach. Thus these naive applications of amplitude estimation algorithm are undesirable.

Motivated by these observations, we propose a novel method that retains the $\mathcal{O}(1/\epsilon)$ scaling (where ϵ is the error in the energy per site) while using at most logarithmic number of qubits (in L). We remark that prior works have already advocated using variants/improvements of the amplitude estimation algorithm [37] for variational energy calculations [62–65]. However, it is a priori not obvious if these approaches can be used to achieve the desired $\mathcal{O}(1/\epsilon)$ accuracy using at most logarithmic number of qubits (in L), because the original DMERA circuit

[Section II] is defined on L^d qubits.

Without going into the details of our approach, let us first provide an executive summary on its complexity. An important parameter that determines the complexity is the *period* of the DMERA circuit within each scale. For example, for $r = 2$, within each scale, if all the gates are chosen identically, shifting all the qubits by 2 (in the unit of the underlying lattice) yields the identical circuit. In this case, the period is 2. More generally, the period can be increased by (i) choosing a larger r or (ii) choosing the gates in a non-uniform way [Fig. 12]. Thus the period depends on two parameters: (i) the quantum circuit architecture and (ii) the pattern of the gates being chosen.

Throughout this Section, for simplicity we shall assume the same period for different directions, and denote the period as $\mathcal{T}$. The overall complexity of the algorithm can be summarized as follows.

- Gate count: $\mathcal{O}\left(\frac{\max(D,\mathcal{T})^{d+1}\log N}{\epsilon}\right)$
- Number of qubits: $\Theta(\max(D,\mathcal{T})^d \log N)$.

Recall that $\mathcal{T}$ is a period, $N = 2^{nd}$ is the system size, D is the depth of the DMERA per unit layer, and ϵ is the relative precision in the energy. Thus, provided that the period $\mathcal{T}$ is sufficiently small compared to D, the extra cost in the gate count is negligible. Number of qubits required is increased, but only slightly, by a factor logarithmic in N.

We remark that $\mathcal{T}$ being a small number is a reasonable assumption. In many existing examples, the gates are chosen in a translationally invariant way [45, 56, 58, 59], which immediately implies $\mathcal{T} = \mathcal{O}(1)$. Therefore, under such assumption one can achieve a gate count of $\mathcal{O}(1/\epsilon)$ and a qubit count of at most $\mathcal{O}(\log N)$.

The rest of this Section is organized as follows. We first explain in Section VI A the amplitude estimation algorithm [37], which is the algorithm that our algorithm builds upon. We consider a few ways in which this algorithm can be applied to our setup in a "black box" fashion, and explain why these approaches are inadequate for our purpose. In Section VI B, we emphasize a special structure of DMERA which we can utilize to obtain the desired complexity.

A. Amplitude Estimation algorithm

In the amplitude estimation algorithm [37], one assumes access to a unitary U that prepares some state $|\psi\rangle$ from a simple product state,. e.g., $|0\ldots0\rangle$ and a reflection $R = I - 2P$, where P is a projection. The cost of this algorithm can be quantified in terms of the number of invocation of the Grover-type operator:

$$\mathcal{Q} = UR_0U^\dagger R, \tag{18}$$

where R is the reflection about the $|0\ldots0\rangle$ state. This algorithm estimates $\langle\psi|P|\psi\rangle$ with an additive precision of ϵ, using $\mathcal{O}(1/\epsilon)$ invocations of $\mathcal{Q}$.

It is well-known that the amplitude estimation algorithm can be used to estimate expectation values of certain observables with $\mathcal{O}(1/\epsilon)$ complexity, assuming one has access to a gate of the form of $e^{i\theta O}$ for an observable O that one intends to measure and a set of $\theta \in \mathbb{R}$. Let us remark that if the observable of interest can be decomposed into a linear combination of $\mathcal{O}(1)$ number of multi-qubit Pauli matrices, the fact that this is possible is rather obvious; in this case one can simply choose a reflection as one of the individual Paulis, estimate their expectation values individually, and take a sum.

As we discussed before, a naive application of the amplitude estimation algorithm to DMERA yields the $\mathcal{O}(1/\epsilon)$ complexity, but with caveats. Doing so appears to increase the gate complexity in N (system size) from $\mathcal{O}(\log N)$ to $\mathcal{O}(N \log N)$. Randomly selecting an observable (in the spirit of Section V) and applying amplitude estimation for the set of chosen observables yields a $\mathcal{O}(\log N)$ scaling, but now the scaling in ϵ becomes $\mathcal{O}(1/\epsilon^2)$, unless the variance of the expectation value of the observable (with respect to the choice of the observable happens) is zero. This is unlikely, because even if the gates in the DMERA are chosen in a translationally invariant way, the resulting state may not be translationally invariant.

B. Coherent Sampling: Amplitude Estimation for DMERA

The naive approaches explained above fail to achieve $\mathcal{O}(\epsilon^{-1}\log N)$ gate complexity. Here we propose a method that achieves this goal, referred to as the *coherent sampling* approach. The key idea is to construct a compact circuit that prepares a purification of the uniform mixture of the reduced density matrices that support individual terms of the Hamiltonian. Suppose, for example, that we are interested in estimating the expectation value of a nearest-neighbor Pauli operator P supported on (x, y) and $(x + 1, y)$, averaged uniformly over x, which we denote as $\bar{P}$. To estimate the expectation value of this observable, we will seek to prepare a purification of the reduced density matrix over (x, y) and $(x + 1, y)$, averaged uniformly over both x and y. It is then easy to see that the the expectation value of $\bar{P}$ with respect to the original state is equal to the expectation value of P with respect to the averaged density matrix. Thus, if we can construct a circuit that prepares a purification of the averaged reduced density matrix, we can apply the amplitude estimation algorithm to that circuit to measure expectation value of $\bar{P}$ with $\mathcal{O}(1/\epsilon)$ complexity.

Such a purification can be constructed by noting that within every scale of the DMERA, there are at most $\mathcal{T}^d$ distinct circuits, all related to each other by shifting the qubits in the x- and y-directions. As a simple exam-

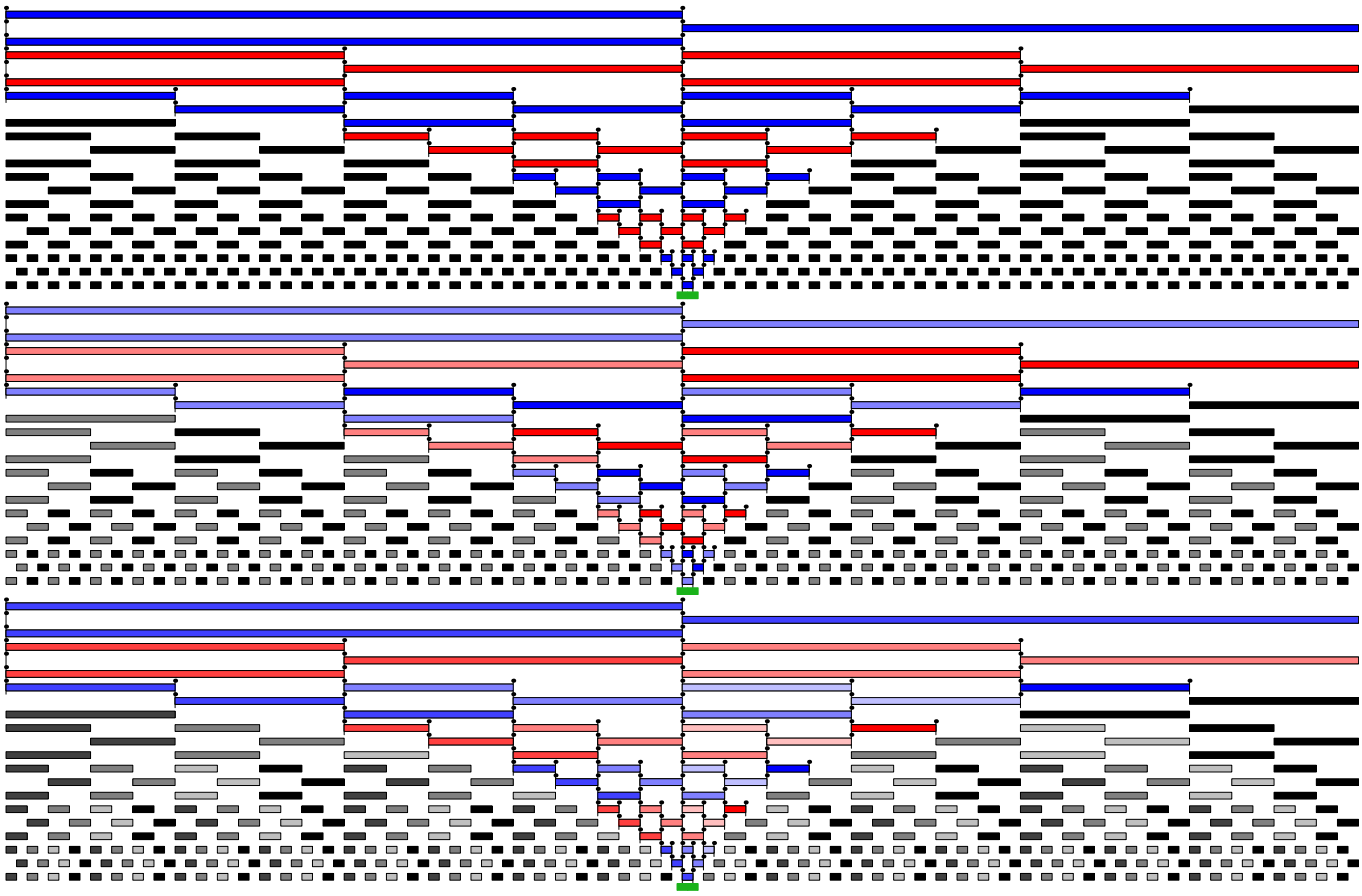


Figure 12. DMERA with a with translationally invariant gates with period 2 (top), 4 (middle) and 8 (bottom). Here we chose $n = 7$ and $D = 3$ for concreteness. Within each layer, the gates with the same level of gradation are chosen to be the same gates. Blue/red gates are the gates that appear in the past causal cone of a nearest-neighbor observable (green).

ple, consider a one-dimensional example, shown in Fig. 6. Here one can see that the past causal cone within a single scale, for all local observables, can be obtained by shifting a sufficiently wide past causal cone and possibly restricting it to a subset of qubits. Therefore, one can obtain the requisite averaged reduced density matrix by applying a random shift as we transition from one scale to another. Importantly, this can be done coherently, as we explain below.

A useful subroutine is SHIFT. This operation shifts the qubits in the past causal cone of the DMERA, either in the x- or y-direction. Let us denote the linear size of the past causal cone as ℓ. Specifically, let SHIFT be a unitary defined as

$$\text{SHIFT} : \bigotimes_{k=0}^{\ell-1} |\varphi_k\rangle_k \to \bigotimes_{k=0}^{\ell-1} |\varphi_{\sigma_1(k)}\rangle_k, \tag{19}$$

where σ_1 is the cyclic permutation, i.e., $\sigma_1(0) = 1, \sigma_1(1) = 2, \ldots \sigma_1(\ell-1) = 0$ and $|\varphi_k\rangle_k$ is a ℓ-qubit state. We can similarly define cSHIFT (an acronym for controlled-SHIFT) as

$$c\text{SHIFT} : |x\rangle \bigotimes_{k=0}^{\ell-1} |\varphi_k\rangle_k \to |x\rangle \bigotimes_{y=0}^{\ell-1} |\varphi_{\sigma_x(k)}\rangle_k, \tag{20}$$

where σ_0 is the identity and $\sigma_x = (\sigma_1)^x$ is a cyclic permutation by $x \in \{0, \ldots, \ell-1\}$. For an implementation of cSHIFT, see Fig. 13.

We note that any cyclic permutation over m elements can be decomposed into at most m SWAPs. In two dimensions, because there are at most ℓ rows/columns in the past causal cone, any power of SHIFT can be realized using at most ℓ^2 SWAPs. Therefore, using the construction in Fig. 13, we conclude the cost of cSHIFT to be at most $\ell^2\lceil \log_2 \mathcal{T} \rceil$ controlled-SWAPs. Each controlled-SWAP can be decomposed into one TOFFOLI and two CNOT gates. Thus, cSHIFT costs at most $\ell^2\lceil \log_2 \mathcal{T} \rceil$ Toffolis and $2\ell^2\lceil \log_2 \mathcal{T} \rceil$ CNOT gates. In particular, it consumes at moat $8\ell^2\lceil \log_2 \mathcal{T} \rceil$ T-gates [60].

Now let us discuss how to use these operations to obtain the averaged density matrix in DMERA. One can do this by essentially randomizing the way in which the output from one scale is connected to the input to the other

scale. In 1D DMERA, for each scale, there are exactly ℓ possible ways in which the output from one scale is connected to the input to the next (lower) scale. The number of ancilla qubits needed is $\lceil \log_2 \mathcal{T} \rceil$ using the standard circuit construction. Similarly, in 2D DMERA, for each scale, there are exactly ℓ^2 possible ways in which the output from one scale is connected to the input to the next (lower) scale. The total ancilla qubit number is $2\lceil \log_2 \mathcal{T} \rceil$ and the gate count being $2\ell^2 \lceil \log_2 \mathcal{T} \rceil$ controlled-SWAPs.

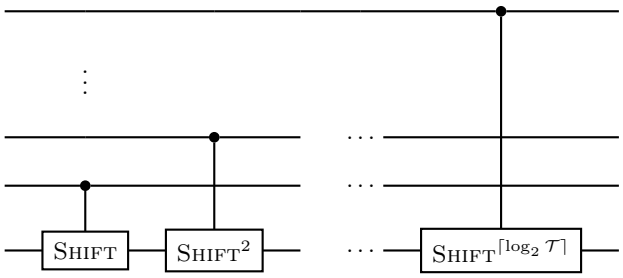


Figure 13. Using cSHIFT to randomize the connection between different scales.

To coherently randomize these shifts, we must prepare the ancillary register in the uniform superposition over $\{0, \ldots, \mathcal{T}-1\} \times \{0, \ldots, \mathcal{T}-1\}$. The cost of this is known to be $\mathcal{O}(\log \mathcal{T})$; see Ref. [9] for details.

C. Resource estimate

In this Section, we perform a more detailed resource estimate of the coherent sampling algorithm. In particular, we focus on estimating the overall T-gate cost and the number of qubits associated to the implementation of $\mathcal{Q}$ [Eq. (18)]. Implementation of $\mathcal{Q}$ consists of three components: cSHIFT, preparation of the uniform superposition state over $\{0, \ldots, \mathcal{T}-1\} \times \mathcal{T}-1\}$, and application of the gates lying in the past causal cone.

Here is a breakdown of the T-gate count. The cSHIFT and the uniform superposition state preparation incurs $\mathcal{O}(\ell^d \log \mathcal{T})$ and $\mathcal{O}(d \log \mathcal{T})$ respectively, as discussed in Section VI B. The number of two-qubit gates lying in the past causal cone can be estimated to be $\mathcal{O}(\ell^{d+1} \log N)$, where $\ell = \Theta(\max(D, \mathcal{T}))$ [Section III, VI B]. Thus the overall cost is $\mathcal{O}(\max(D, \mathcal{T})^{d+1} \log N)$. Recalling that the query complexity(for $\mathcal{Q}$) of the amplitude estimation algorithm [37] is $\mathcal{O}(1/\epsilon)$,we arrive at the overall cost of $\mathcal{O}\left(\max(D, \mathcal{T})^{d+1} \epsilon^{-1} \log N\right)$.[66]

The number of qubits can be estimated as follows. For implementing the gates in the past causal cone, there are $\Theta(\max(D, \mathcal{T})^d \log N)$ qubits. For the cSHIFT and the preparation of uniform superposition state, there are $\Theta(\log \mathcal{T})$ additional qubits needed. Thus the overall qubit count is $\Theta(\max(D, \mathcal{T})^d \log N)$.

VII. CONCLUSION

In this paper, we explored the possibility of using variational ansatzes such as DMERA [20] to study the ground state properties of a system consisting of a large number of fermions interacting with each other. The main strength of this approach is that it requires a relatively smaller number of qubits compared to the approaches based on QPE. For the specific case we studied, consisting of a system of size 64×64, we observed that indeed a large reduction in the number of qubits is possible [Section V].

We also studied the pertinent issues concerning variational approaches, such as the large number of measurements needed [32] and the Barren plateau problem [33]. We found that the measurement problem still persisted even for a simple model (e.g., Fermi-Hubbard model in two spatial dimensions), suggesting that the conventional QPE-based approaches will be more advantageous if a sufficiently large quantum computer is available. The root of this problem lies in the inevitable $\mathcal{O}(1/\epsilon^2)$ scaling for the precision ϵ that is present in any sampling-based approach. Fortunately, we found that it is possible to restore the Heisenberg scaling whilst only incurring an additional logarithmic overhead in the system size [Section VI A].

For the Barren plateau problem, we found that the typical magnitude of the gradient depends greatly on the location of the gate; the deeper it is in the circuit, smaller the gradient is. More precisely, we can organize the DMERA circuit into $n = \log_2 N$ layers, each corresponding to a different length scale [Fig. 6]. The gates appearing the coarsest length scale have a variance decaying exponentially in Dn, yielding a polynomially small number. However, at the finest length scale, the variance decays exponentially in D alone, which is a constant indepent of the system size. At the s'th layer (setting $s = 0$ to be the finest and $s = n$ to be the coarsest scale), the variance decays exponentially with ns. Thus for DMERA, the Barren plateau problem can be avoided as a function of system size.

There are several natural questions to pursue, which we leave for future work. First, a more detailed comparison between the coherent and incoherent sampling approach — taking into account of the constant factors — would be desirable. Second, it will be desirable to develop a theory on how well the approximation error scales with D. The existing studies [25, 26] are limited to non-interacting systems. Would this picture change or remain intact in the presence of interaction? Perhaps experiments involving small-scale quantum computers will be illuminating. Third, understanding the expressiveness of the 2D DMERA remains as an important open problem. In Ref. [67], the authors investigated how well tensor network ansatzes suitable for studying many-body systems in one spatial dimension [20, 30, 68] can approximate ground states of physical interest.[69] Extending such a study to two spatial dimensions is an important

open problem. Lastly, due to a very modest resource requirement reported in Section V, the fault-tolerance overhead needed to make the effect of noise negligible may be much more benign than the requirements for other approaches based on QPE. Quantifying this amount will be another interesting direction to pursue.

ACKNOWLEDGEMENT

This work was supported in part by funding from Lockheed Martin.

[1] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, Quantum computational chemistry, Reviews of Modern Physics **92**, 015003 (2020).

[2] R. P. Feynman, Simulating Physics with Computers, Int. J. Theor. Phys. **21**, 467 (1982).

[3] S. Lloyd, Universal quantum simulators, Science **273**, 1073 (1996).

[4] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, Elucidating reaction mechanisms on quantum computers, Proceedings of the National Academy of Sciences **114**, 7555 (2017).

[5] G. H. Low and I. L. Chuang, Hamiltonian Simulation by Qubitization, Quantum **3**, 163 (2019).

[6] D. W. Berry, C. Gidney, M. Motta, J. R. McClean, and R. Babbush, Qubitization of Arbitrary Basis Quantum Chemistry Leveraging Sparsity and Low Rank Factorization, Quantum **3**, 208 (2019).

[7] J. Lee, D. W. Berry, C. Gidney, W. J. Huggins, J. R. McClean, N. Wiebe, and R. Babbush, Even More Efficient Quantum Computations of Chemistry Through Tensor Hypercontraction, PRX Quantum **2**, 030305 (2021).

[8] V. von Burg, G. H. Low, T. Häner, D. S. Steiger, M. Reiher, M. Roetteler, and M. Troyer, Quantum computing enhanced computational catalysis, Physical Review Research **3**, 033055 (2021), arXiv:2007.14460 [physics, physics:quant-ph].

[9] R. Babbush, C. Gidney, D. W. Berry, N. Wiebe, J. McClean, A. Paler, A. Fowler, and H. Neven, Encoding Electronic Spectra in Quantum Circuits with Linear T Complexity, Physical Review X **8**, 041015 (2018).

[10] R. Babbush, D. W. Berry, J. R. McClean, and H. Neven, Quantum simulation of chemistry with sublinear scaling in basis size, npj Quantum Information **5**, 92 (2019).

[11] Y. Su, D. W. Berry, N. Wiebe, N. Rubin, and R. Babbush, Fault-tolerant quantum simulations of chemistry in first quantization, PRX Quantum **2**, 040332 (2021).

[12] A. Delgado, P. A. M. Casares, R. dos Reis, M. S. Zini, R. Campos, N. Cruz-Hernández, A.-C. Voigt, A. Lowe, S. Jahangiri, M. A. Martin-Delgado, J. E. Mueller, and J. M. Arrazola, Simulating key properties of lithium-ion batteries with a fault-tolerant quantum computer, Phys. Rev. A **106**, 032428 (2022).

[13] M. Shokrian Zini, A. Delgado, R. dos Reis, P. A. Moreno Casares, J. E. Mueller, A.-C. Voigt, and J. M. Arrazola, Quantum simulation of battery materials using ionic pseudopotentials, Quantum **7**, 1049 (2023).

[14] M. Fiebig, T. Lottermoser, D. Meier, and M. Trassin, The evolution of multiferroics, Nature Reviews Materials **1**, 1 (2016).

[15] N. A. Spaldin and R. Ramesh, Advances in magnetoelectric multiferroics, Nature Materials **18**, 203 (2019).

[16] J. Wang, J. B. Neaton, H. Zheng, V. Nagarajan, S. B. Ogale, B. Liu, D. Viehland, V. Vaithyanathan, D. G. Schlom, U. V. Waghmare, N. A. Spaldin, K. M. Rabe, M. Wuttig, and R. Ramesh, Epitaxial bifeo3 multiferroic thin film heterostructures, Science **299**, 1719 (2003).

[17] G. Catalan and J. F. Scott, Physics and Applications of Bismuth Ferrite, Advanced Materials **21**, 2463 (2009).

[18] Y.-H. Chu, M. P. Cruz, C.-H. Yang, L. W. Martin, P.-L. Yang, J.-X. Zhang, K. Lee, P. Yu, L.-Q. Chen, and R. Ramesh, Domain Control in Multiferroic BiFeO3 through Substrate Vicinality, Advanced Materials **19**, 2662 (2007).

[19] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a quantum processor, Nature Communications **5**, 10.1038/ncomms5213 (2014), 1304.3061v1.

[20] I. H. Kim, Noise-resilient preparation of quantum many-body ground states, arXiv:1703.00032 (2017), arXiv:1703.00032 [quant-ph].

[21] G. Vidal, Class of quantum many-body states that can be efficiently simulated, Phys. Rev. Lett. **101**, 110501 (2008).

[22] G. Evenbly and G. Vidal, Tensor Network States and Geometry, Journal of Statistical Physics **145**, 891 (2011), arXiv:1106.1082 [quant-ph].

[23] M. Aguado and G. Vidal, Entanglement renormalization and topological order, Phys. Rev. Lett. **100**, 070404 (2008).

[24] R. König, B. W. Reichardt, and G. Vidal, Exact entanglement renormalization for string-net models, Phys. Rev. B **79**, 195123 (2009).

[25] J. Haegeman, B. Swingle, M. Walter, J. Cotler, G. Evenbly, and V. B. Scholz, Rigorous free fermion entanglement renormalization from wavelet theory, ArXiv e-prints (2017), arXiv:1707.06243 [quant-ph].

[26] G. Evenbly and S. R. White, Entanglement renormalization and wavelets, Phys. Rev. Lett. **116**, 140403 (2016).

[27] S. R. White, Density matrix formulation for quantum renormalization groups, Phys. Rev. Lett. **69**, 2863 (1992).

[28] F. Verstraete, M. M. Wolf, D. Perez-Garcia, and J. I. Cirac, Criticality, the area law, and the computational power of projected entangled pair states, Phys. Rev. Lett. **96**, 220601 (2006).

[29] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, Annals of Physics **349**, 117 (2014).

[30] I. H. Kim, Holographic quantum simulation, arXiv:1702.02093 (2017), arXiv:1702.02093 [quant-ph].

[31] I. H. Kim and B. Swingle, Robust entanglement renormalization on a noisy quantum computer, (2017),

arXiv:1711.07500.
[32] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015), arXiv:1507.08969.
[33] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nature Communications **9**, 4812 (2018).
[34] T. Barthel and Q. Miao, Absence of barren plateaus and scaling of gradients in the energy optimization of isometric tensor network states (2023), arXiv:2304.00161 [quant-ph].
[35] I. D. Kivlichan, C. Gidney, D. W. Berry, N. Wiebe, J. McClean, W. Sun, Z. Jiang, N. Rubin, A. Fowler, A. Aspuru-Guzik, H. Neven, and R. Babbush, Improved Fault-Tolerant Quantum Simulation of Condensed-Phase Correlated Electrons via Trotterization, Quantum **4**, 296 (2020).
[36] E. T. Campbell, Early fault-tolerant simulations of the hubbard model, Quantum Science and Technology **7**, 015007 (2021).
[37] G. Brassard, P. Hoyer, M. Mosca, and A. Tapp, Quantum Amplitude Amplification and Estimation (2002) pp. 53–74, arXiv:quant-ph/0005055.
[38] P. Corboz and G. Vidal, Fermionic multiscale entanglement renormalization ansatz, Physical Review B **80**, 165129 (2009).
[39] I. Arad and Z. Landau, Quantum computation and the evaluation of tensor networks (2010), arXiv:0805.0040 [quant-ph].
[40] P. Jordan and E. Wigner, Über das Paulische Äquivalenzverbot, Zeitschrift für Physik **47**, 631 (1928).
[41] G. Evenbly and G. Vidal, Algorithms for entanglement renormalization, Phys. Rev. B **79**, 10.1103/physrevb.79.144108 (2009).
[42] T. Grover, Entanglement monotonicity and the stability of gauge theories in three spacetime dimensions, Phys. Rev. Lett. **112**, 151601 (2014).
[43] J. G. Bednorz and K. A. Müller, Possible high-tc superconductivity in the ba-la-cu-o system, Zeitschrift für Physik B Condensed Matter **64**, 189 (1986).
[44] L. Cincio, J. Dziarmaga, and M. M. Rams, Multiscale entanglement renormalization ansatz in two dimensions: Quantum ising model, Phys. Rev. Lett. **100**, 240603 (2008).
[45] G. Evenbly and G. Vidal, Entanglement renormalization in noninteracting fermionic systems, Phys. Rev. B **81**, 235102 (2010).
[46] G. Evenbly and G. Vidal, Tensor Network Renormalization Yields the Multiscale Entanglement Renormalization Ansatz, Physical Review Letters **115**, 200401 (2015), arXiv:1502.05385 [cond-mat.str-el].
[47] Changing these choices will not alter the conclusions of this paper significantly, although the precise numbers we estimate later are likely to depend on those choices.
[48] V. Giovannetti, S. Montangero, and R. Fazio, Quantum multiscale entanglement renormalization ansatz channels, Phys. Rev. Lett. **101**, 180503 (2008).
[49] D. B. Rubin, The bayesian bootstrap, The annals of statistics , 130 (1981).
[50] G. Vidal and C. M. Dawson, Universal quantum circuit for two-qubit transformations with three controlled-not gates, Phys. Rev. A **69**, 010301 (2004).
[51] A. Bocharov, M. Roetteler, and K. M. Svore, Efficient synthesis of universal repeat-until-success quantum circuits, Phys. Rev. Lett. **114**, 080502 (2015).
[52] V. Kliuchnikov, D. Maslov, and M. Mosca, Asymptotically optimal approximation of single qubit unitaries by clifford and t circuits using a constant number of ancillary qubits, Phys. Rev. Lett. **110**, 190502 (2013).
[53] D. Gosset, V. Kliuchnikov, M. Mosca, and V. Russo, An algorithm for the t-count, Quantum Info. Comput. **14**, 1261–1276 (2014).
[54] A. Bocharov, M. Roetteler, and K. M. Svore, Efficient synthesis of probabilistic quantum circuits with fallback, Phys. Rev. A **91**, 052317 (2015).
[55] N. J. Ross and P. Selinger, Optimal ancilla-free clifford+t approximation of z-rotations, Quantum Info. Comput. **16**, 901–953 (2016).
[56] C. T. Olund, M. Block, S. Gazit, J. McGreevy, and N. Y. Yao, Adiabatic ground state preparation in an expanding lattice, Physical Review B **101**, 155152 (2020).
[57] S. Anand, J. Hauschild, Y. Zhang, A. C. Potter, and M. P. Zaletel, Holographic quantum simulation of entanglement renormalization circuits (2022), arXiv:2203.00886 [quant-ph].
[58] Q. Miao and T. Barthel, Quantum-classical eigensolver using multiscale entanglement renormalization, Phys. Rev. Res. **5**, 033141 (2023).
[59] Q. Miao and T. Barthel, Convergence and quantum advantage of trotterized mera for strongly-correlated systems (2023), arXiv:2303.08910 [quant-ph].
[60] C. Gidney, Halving the cost of quantum addition, Quantum **2**, 74 (2018).
[61] J. P. F. LeBlanc, A. E. Antipov, F. Becca, I. W. Bulik, G. K.-L. Chan, C.-M. Chung, Y. Deng, M. Ferrero, T. M. Henderson, C. A. Jiménez-Hoyos, E. Kozik, X.-W. Liu, A. J. Millis, N. V. Prokof'ev, M. Qin, G. E. Scuseria, H. Shi, B. V. Svistunov, L. F. Tocchio, I. S. Tupitsyn, S. R. White, S. Zhang, B.-X. Zheng, Z. Zhu, and E. Gull (Simons Collaboration on the Many-Electron Problem), Solutions of the two-dimensional hubbard model: Benchmarks and results from a wide range of numerical algorithms, Phys. Rev. X **5**, 041041 (2015).
[62] D. Wang, O. Higgott, and S. Brierley, Accelerated variational quantum eigensolver, Phys. Rev. Lett. **122**, 140504 (2019).
[63] G. Wang, D. E. Koh, P. D. Johnson, and Y. Cao, Minimizing estimation runtime on noisy quantum computers, PRX Quantum **2**, 010346 (2021).
[64] R. Zhang, G. Wang, and P. Johnson, Computing Ground State Properties with Early Fault-Tolerant Quantum Computers, Quantum **6**, 761 (2022).
[65] A. Katabarwa, A. Kunitsa, B. Peropadre, and P. Johnson, Reducing runtime and error in vqe using deeper and noisier quantum circuits (2021), arXiv:2110.10664 [quant-ph].
[66] There are additional gates used in the amplitude estimation algorithm, such as the implementation of the reflection R. However, the associated cost is negligible in comparison.
[67] R. Haghshenas, J. Gray, A. C. Potter, and G. K.-L. Chan, Variational power of quantum circuit tensor networks, Phys. Rev. X **12**, 011047 (2022).
[68] M. Foss-Feig, D. Hayes, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, J. M. Pino, and A. C. Potter, Holographic quantum algorithms for simulating corre-

lated spin systems, Phys. Rev. Res. **3**, 033002 (2021).
[69] We remark, however, that Ref. [67] also studies how well certain two-dimensional systems are approximated by a one-dimensional ansatz such as MPS.
[70] For instance, even if the same gate is used twice, those two different instantiations of the same gate will be associated with two different identification numbers.
[71] The main bottleneck in practice is Line 7 of Algorithm 2. A judiciously choice of data structure seems to be necessary to make this approach practical.
[72] S. B. Bravyi and A. Y. Kitaev, Fermionic Quantum Computation, Ann. Phys. **298**, 210 (2002), arXiv:0003137 [quant-ph].
[73] P. Corboz, G. Evenbly, F. Verstraete, and G. Vidal, Simulation of interacting fermions with entanglement renormalization, Phys. Rev. A **81**, 010303 (2010), arXiv:0904.4151 [cond-mat.str-el].
[74] F. Verstraete and J. I. Cirac, Mapping local hamiltonians of fermions to local hamiltonians of spins, J. Stat. Mech. **0509**, P09012 (2005), cond-mat/0508353v3.
[75] J. D. Whitfield, V. Havlíček, and M. Troyer, Local spin operators for fermion simulations, Physical Review A **94**, 030301 (2016).
[76] M. Steudtner and S. Wehner, Quantum codes for quantum simulation of fermions on a square lattice of qubits, Physical Review A **99**, 022308 (2019).
[77] K. Setia, S. Bravyi, A. Mezzacapo, and J. D. Whitfield, Superfast encodings for fermionic quantum simulation, Physical Review Research **1**, 033033 (2019).
[78] Z. Jiang, J. McClean, R. Babbush, and H. Neven, Majorana Loop Stabilizer Codes for Error Mitigation in Fermionic Quantum Simulations, Physical Review Applied **12**, 064041 (2019).
[79] C. Derby, J. Klassen, J. Bausch, and T. Cubitt, Compact fermion to qubit mappings, Physical Review B **104**, 035118 (2021).
[80] Y.-A. Chen and Y. Xu, Equivalence between fermion-to-qubit mappings in two spatial dimensions (2022), arXiv:2201.05153 [cond-mat, physics:math-ph, physics:quant-ph].

Appendix A: Compilation approaches

In this Appendix, we introduce compilation algorithms used in obtaining Fig. 11. Specifically, we introduce two algorithms, referred to as Algorithm 1 and Algorithm 2. Algorithm 1 has the advantage of being relatively fast, but leads to suboptimal number of qubits. In contrast, Algorithm 2 leads to a reduced qubit cost but is slower than Algorithm 1.

Throughout this Appendix, we shall assume that a circuit is described by the following data structure. A `gate` is an object with the property called `support`, which specifies the set of qubits the `gate` acts on. A `layer` is a collection of `gate`s, each of which have disjoint `support`s. Intuitively, a layer can be thought as a unit-depth quantum circuit. A `circuit` is a sequence of `layer`s. To summarize, a `circuit` is a sequence of `layer`s, each of which consists of `gate`s that commute with each other.

The goal of the compilation algorithms is to assign the qubits appearing in the `support` of each `gate` to the set of qubits in the quantum computer. The algorithms discussed below achieve this goal, with additional gates that reset the state of the qubit. The cost of these extra gates are negligible and they are useful for reducing the number of qubits.

We shall assume access to the following simple functions, whose implementation is straightforward. `Qubits` is a function that, applied to an object, returns a set of qubits that the object nontrivially acts on. for instance, applied to a `layer` object, it returns the union of the `support`s of the `gate`s in the `layer`. Applied to a `circuit` layer, it returns the union of `Qubits(layer)` for all the `layer`s within the `circuit`. `Length` is a function that, applied to an object discussed above, returns the number of elements of the object. For instance, applied to a `circuit` object, this function returns the number of `layer`s within the `circuit` object. Applied to a `layer`, it returns the number of `gate` objects within the `layer`. Applied to the set of qubits, it returns the cardinality of the set.

The basic idea behind Algorithm 1 is to iterate through each layer in the following way. Given a layer, we can divide the circuit into two pieces. One of them is the "past," which we denote as `Past(layer, circuit)`. This is a sequence of `layer`s within the `circuit` up to the `layer` under consideration. The other is the "future," which we denote as `Future(layer, circuit)`. This is the sequence of `layer`s appearing after the `layer` under consideration, within the `circuit`. Note that the qubits in the past not appearing in the future need not be kept, because they are never used afterwards. Therefore, we can simply reset these qubits, freeing them up for future use; this corresponds to Line 8 of Algorithm 1.

The pseudocode description of Algorithm 1 uses two additional operations, denoted as `Assign` and `Remove`. Both these operations implicitly assume that there is a globally accessible lookup table from the set of qubits appearing in the circuit description to the set of qubits in the quantum computer. The set of qubits available on a quantum computer is assumed to be an ordered list, and the `Assign` makes/updates the assignment in the following greedy way. If the assignment for a qubit in the circuit description is already specified in the lookup table, `Assign` simply applies this lookup table to the `support`s of the `gate`s within each `layer`. If not, `Assign` updates the assignment to the first qubit (on the quantum computer) that is not yet assigned and then repeat the aforementioned procedure. The `Remove` operation removes the assignment in the lookup table for the specified set of qubitss (in the circuit description).

Algorithm 1 High-level description of Algorithm 1.

```
Input: Circ
Output: CircNew
1: ℓ ← Length(Circ)
2: CircNew ← []
3: for i in range(ℓ) do
4:     CircNew.append(Assign(Circ[i]))
5:     PastCirc ← Circ[:i]
6:     FutureCirc ← Circ[i:]
7:     RemoveQ ← Qubits(FutureCirc)-Qubits(PastCirc)
8:     CircNew.append(Reset(Assign(RemoveQ)))
9:     Remove(RemoveQ)
return CircNew
```

Algorithm 2 High-level description of Algorithm 2.

```
Input: Circ
Output: CircNew
1: CircNew ← []
2: while Circ is non-empty do
3:     Qs ← Qubits(Circ)
4:     PCCSizes ← { Q:[Length(Qubits(PCC(Q, Circ)))
   for Q in Qs]}
5:     Qmin ← argmin(PCCSizes)
6:     CircNew.append(Assign(PCC(Qmin, Circ)))
7:     Circ ← Circ - PCC(Qmin, Circ)
8:     CircNew.append(Reset(Assign(Qmin)))
9:     Remove(Qmin)
return CircNew
```

To explain the basic idea behind Algorithm 2 it will be useful to first start with a motivating observation. Without loss of generality, let `q` be a qubit appearing in the circuit and let `PCC(q, Circ)` be the past causal cone of `q` with respect to the `circuit Circ`. For any gate in `Circ` not appearing in the past causal cone, it is possible to move that gate past all the gates appearing in the past causal cone. The proof of this claim is straightforward. If the gate appears after the past causal cone, the claim is true by construction. Otherwise, if a gate lies between gates appearing in the past causal cone and does not commute with at least one gate in the past causal cone appearing afterwards, that gate must necessarily lie in the past causal cone; this is a contradiction. Therefore, for any `q`, the unitary generated by `Circ` can be decomposed into the unitary generated by `PCC(q, Circ)` followed by other gates.

The basic idea of Algorithm 2 is to compile `Circ` by decomposing it into a past causal cone of a qubit for which the number of qubits in the past causal cone is minimized, followed by the remaining set of gates. The main virtue of this construction is that, once the past causal cone of the qubit is applied, that qubit can be removed, akin to the process described in Algorithm 1. Then the same procedure can be applied to the remainder of the circuit repeatedly.

The pseudocode description of Algorithm 2 makes use of the operations used in Algorithm 1 such as `Assign` and `Remove`. However, Line 4 and 7 requires more explanation. In Line 4, `PCCSizes` is a dictionary data structure with the keys corresponding to the set of qubits in the circuit and the values being the number of qubits in the past causal cone of that qubit. In Line 7, the subtraction of `PCC(Qmin, Circ)` from `Circ` means simply removing the gates in `PCC(Qmin, Circ)` from `Circ`. To be more precise, view `Circ` as a sequence of `gates`, each with a unique identification number.[70] Line 7 means removing the set of gates — labeled in terms of their identification number — in the past causal cone from `Circ`.

Let us make a side remark: that in practice Algorithm 2 is substantially slower than Algorithm 1. Due to this reason, in making Fig. 11 we simply estimated the number of qubits needed on a quantum computer, instead of running the algorithm in Algorithm 2.[71]

Now that we have explained the two algorithms, it is interesting to discuss which of them is of more practical use. As explained above, Algorithm 1 is faster but leads to a suboptimal number of qubits. On the other hand, Algorithm 2 is slower but leads to a smaller number of qubits. We believe, at least in the foreseeable future, it is desirable to use Algorithm 2. Qubits, especially logical qubits, are likely to remain a scarce resource, and as such, it is always desirable to minimize that number. Moreover, the compilation can be done offline prior to the quantum computation. Since classical computation resources are significantly cheaper than the quantum ones, a large compilation cost appears to be an acceptable cost.

Appendix B: Compiling Fermionic DMERA

In this Appendix, we introduce a method to compile a fermionic variant of DMERA, fDMERA for short. This variant of DMERA has the same circuit architecture as DMERA, but assumes that the global Hilbert space and the gates used are fermionic. Fermionic gate refers to a gate acting on a set of (second-quantized) fermions which preserve the fermion parity. A universal set of fermionic gates was first introduced by Bravyi and Kitaev [72], summarized below:

$$\begin{aligned} &\exp\Big(\theta(e^{i\varphi}a_i^\dagger a_j + h.c.)\Big), \\ &\exp\Big(\theta(e^{i\varphi}a_i^\dagger a_j^\dagger + h.c.)\Big), \\ &\exp(\theta n_i n_j), \end{aligned} \tag{B1}$$

where $\theta, \varphi \in \mathbb{R}$ are real parameters, $a_i^\dagger$ and a_i are fermion creation and annihilation operators respectively, and $n_i = a_i^\dagger a_i$. Fermionic DMERA can be simply thought as a quantum circuit with a hierarchical structure (see Section II), consisting of fermionic gates. These circuits was studied both numerically [73] and analytically [25].

Since quantum computers that are likely going to be available will consist of qubits, to use fermionic ansatzes a conversion from a fermionic circuit to a circuit acting on a set of qubits is needed. One could, in principle, use the standard Jordan-Wigner(JW) transformation [40]. However, JW transformation maps few-body fermionic observables to a many-body spin observable, which is undesirable. There are other alternative approaches which preserve locality [72, 74–80]. Upon applying these methods, one would be able to convert the gates in Eq. (B1) to unitaries acting on $\mathcal{O}(1)$ qubits. However, such a mapping often increases the number of qubits needed.

Motivated from these facts, we provide an alternative approach that can minimize the number of qubits needed. The basic idea is to first compute the past causal cone *without converting the fermions to qubits*. This ends up being a fermionic circuit acting on a smaller set of fermions. By applying the JW transformation to this reduced circuit, we can convert this circuit to a qubit circuit without incurring any extra qubit overhead.

Recall that in Appendix A we introduced methods to reduce the number of qubits needed, by keeping track of qubits that are no longer in use and resetting them, freeing them up. Analogous tricks can be applied to fermions as well. However, what is not so clear is what it means to reset a fermion when we convert that process into a physical operation acting on a set of qubits, via the JW transformation. This is what we intend to discuss primarily in this Appendix.

Without loss of generality, suppose we would like to reset the k'th fermion to be in the unoccupied state. In the fermionic occupation basis, the operation that we intend to perform is

$$|x\rangle_f \to |x_k\rangle_f, \tag{B2}$$

where $|x\rangle_f$ is a basis state of the Fock basis represented by a bit-string x and x_k can be obtained by setting the k'th bit of x to 0. This can be done by (i) measuring the k'th fermion occupation number of and (ii) if the occupation number is 1, apply the annihilation operator a_k.

Under the JW transformation, the occupation number n_k is related to the Pauli operators acting on qubits as $Z_k = 2n_k - I$. The annihilation operator a_k in terms of the Pauli operators is

$$\left(\bigotimes_{j=1}^{k-1} Z_k\right) \bigotimes \frac{X_k - iY_k}{2}, \tag{B3}$$

whose action on the state in which the k'th qubit is set to $Z_k = +1$ is equivalent to the action of the following operator

$$\left(\bigotimes_{j=1}^{k-1} Z_k\right) \bigotimes X_k. \tag{B4}$$

Thus, to reset the k'th fermion to the unoccupied state, it suffices to measure Z_k and apply $\left(\bigotimes_{j=1}^{k-1} Z_k\right) \bigotimes X_k$ only if $Z_k = +1$. Resetting the fermion to the occupied state can be performed in a similar way, by instead using the JW transformation of the fermionic creation operator.