Quantum computers offer computational power fundamentally different from classical computers. A universal computer may solve classically intractable problems within areas ranging from many-body physics to quantum chemistry [1]. There has been impressive experimental progress in developing quantum computers based on superconducting qubits [2], trapped ions [3], and neutral atoms [4]. The need for quantum error-correction to ensure fault-tolerant computation remains a daunting challenge but noisy intermediate-scale quantum (NISQ) devices are expected to be available in the near future [5]. These are devices containing a few hundred qubits with small error rates but without error-correction. An outstanding question is what kind of computations such devices may facilitate.

Algorithms designed for NISQ devices should run on a moderate number of qubits and be resilient to noise. The specific hardware may also pose further restrictions regarding the connectivity of the device in the sense that not all qubits can interact directly with each other [2, 4]. Promising frameworks that fulfill these conditions are the quantum approximate optimization algorithm (QAOA) [6] and the quantum variational eigensolver (VQE) [7, 8]. In these frameworks, the task of the quantum computer is roughly speaking to compute the expectation value of local Hamiltonians on some many-body quantum state. Recent work has characterized a number of conditions for which this can be done in a noise-robust way [9, 10]. This is of paramount importance for implementation with NISQ devices and Refs. [9, 10] thus identify a number of promising near-term applications of quantum computing. Due to the limited resources of NISQ devices, it is, however, also very important to run such algorithms as efficiently as possible in terms of circuit size and number of qubits.

In this letter, we show how to efficiently compute key physical properties of quantum matter on NISQ devices. We upper bound the circuit size and number of qubits necessary to estimate the expectation values of local observables of a many-body quantum state. Importantly, these bounds significantly decrease the resource requirements compared to previous works. Specifically, we are able to show that the energy of a many-body quantum state can be estimated with a constant-sized quantum circuit under similar assumptions as in Refs. [9, 10].

We adopt a general framework for preparing many-body quantum states akin to sequentially generated [11–13] or finitely correlated states [14]. This framework enables us to control the size of the so-called past causal cone [15–17] of local observables. Combined with the notion of mixing rate of local observables under the circuit [9, 10] we determine after how many layers of the circuit, the expectation values of local observables stabilize. To estimate these expectation values, it suffices to implement the potentially small subset of the circuit under which they stabilize instead of producing the entire many-body state. Consequently, the necessary number of qubits and quantum gates can be reduced significantly from scaling with the size of the many-body state to even a constant number.

We consider a framework for producing quantum many-body states that is composed of three basic operations, which are iterated T times as shown in Fig. 1. The first operation is adding new qubits to the existing system. The second operation consists of letting them interact with each other and a bath via a constant depth circuit. As a third and final operation, some of the existing qubits may be discarded.

More specifically, we will start with a system S0 consisting of n0 qubits initialized in some fixed state ρ0 and a bath system B consisting of nB qubits initialized in some fixed state ρB. At each iteration t, we introduce new subsystems St with nt qubits and ancillary states At with αt qubits, all initialized in some fixed quantum state. These new subsystems then interact with the existing ones following an interaction scheme that characterizes how the qubits are connected. Finally, the ancillary system is discarded, which concludes the iteration. The procedure is iterated for a total of T iterations to produce the entire many-body state.
channels of the form $\Phi_t = D_t \circ U_t \circ A_t$. Here $A_t$ adds the new subsystems and auxiliary qubits, $U_t$ is a unitary channel that consists of $D$ two-qubit gates for each edge in $G_t$, and $D_t$ traces out the ancillary systems [18].

Our framework captures a number of widely studied tensor networks states as illustrated in the two examples below.

**Example 1** (Matrix product states and higher dimensional versions). Matrix product states (MPS) have previously been studied in a sequential interaction picture [19, 20] and adapt naturally to our framework. The initial system $S_0$ consists of one qubit and the dimension of the bath gives the bond dimension. At each iteration $t$, we add a system consisting of one qubit $S_t$ to the system. The graph $G_t$ only has edges between the qubits in the bath and the newest added qubit $S_t$. Note that we are considering a proper subset of MPS since we restrict to unitaries implementable with $D$ two-qubit gates for each edge. Arbitrary MPS of a given bond dimension would require arbitrary unitaries. It is straightforward to generalize such sequentially generated states by considering the case in which a bath interacts with a subsystem of dimension $d$ rather than a single qubit at each iteration [9, 12].

**Example 2** (Deep multiscale entanglement renormalization ansatz). The DMERA, introduced in [10], is a variation of the MERA [16] that is tailored for NISQ devices. In our framework, the initial system $S_0$ consists of one qubit and there is no bath. We then define the graphs $G_t$ recursively: at each iteration we add one qubit in between every existing qubit and only nearest neighbors are allowed to interact, resulting in a tree structure.

An important property of our framework is that it allows to bound the number of qubits that can influence the value of a local observable, a number referred to as the causal cone of an observable [15, 16]. The growth of the causal cone depends on the geometry of the graph $G_t$. To see this, we consider an observable $O_T$ on qubits in $G_T$. Note that

$$\text{tr}(\rho O_T) = \text{tr}\left(\left[\Phi_{[0,T]}^* \otimes 1_B\right] \rho \otimes \rho_B\right)$$

by the definition of $\rho$, where $\Phi_{[0,T]}^* = \Phi_T^* \circ \Phi_{T-1}^* \circ \cdots \circ \Phi_0^*$. Here $\Phi_T^*$ is the evolution in the Heisenberg picture.

We wish to keep track of the size of the radius of the support of a local observable $O_T$ on the final state. Going back to the $t$th iteration, we denote the observable $O_t = \Phi_{[0,T]}^*(O_T \otimes 1_B)$. Let $R(O_t)$ be the radius of the smallest ball in $G_t$ containing the support of $O_t$. That is, $O_t$ only differs from the identity on qubits that are at most $2R(O_t)$ edges away in the graph $G_t$. To analyse the growth of the support and its past causal cone we consider the action of $\Phi_T^* = A_T^* \circ U_T^* \circ D_T^*$. First, $D_T^*$ acts by tensoring the identity operator on the auxiliary qubits, not increasing the support. In the next step,
$U_T^*$ increases the support. As $OT$ has radius $R(OT)$, it will be mapped to an observable with radius at most $R(OT) + D$ by $U_T$ according to the locality assumptions of $U_T^*$. The map $A_T^*$ will then map this observable to an observable $OT_{-1}$ supported on qubits that correspond to vertices in $G_{T-1}$, as it traces out all the qubits added at iteration $T$. This can potentially decrease the support of the observable, which is the case in DMERA. Given the graphs $G_i$ and a constant $D$, it is straightforward to keep track of the support of the observable and the past causal cone with the above procedure. This allows us to find the maximum number of unitaries ($N_U(t,r)$) and qubits ($N_Q(t,r)$) in the past causal cone of an observable with radius of support of $r$ on the final state going back to iteration $t$. Note that $N_Q(t,r)$ keeps track of the total number of qubits necessary to implement the past causal cone, and, thus also includes those that were discarded at a previous step. In principle these could be recycled at a later stage of the computation, but we do not analyse this scenario here.

So far we have devised a way of keeping track of the unitaries in the past causal cone of local observables. We are, however also interested in quantifying how much each iteration of the past causal cone contributes to the expectation value. In case the expectation value of the observable stabilizes after a couple of iterations, we can find smaller quantum circuits than the entire causal cone that will approximate the desired expectation value. In other words, only the last $T − t$ steps of the circuit are necessary to approximately compute the expectation value of $OT$ up to an error of $2\delta(t,r)||OT||_\infty$. Note also that the expectation value is independent of the initial state $\rho'$, which we furthermore may restrict to the qubits that are in the support of $O_t$ because only unitaries on these qubits contribute to the expectation value of $OT$. As discussed above, we may further reduce the size of the circuit that needs to be implemented by restricting to the past causal cone. Combining these two observations leads to the formal statement of our main result:

**Theorem.** Let $OT$ be an observable supported in a ball of radius $r$ and $\rho'$ be a state on the qubits that are in the support of $O_t$. It is possible to compute $tr(\rho_{OT})$ up to an additive error $2\delta(t,r)$ by implementing a circuit consisting of $N_U(t,r)$ two-qubit gates on $N_Q(t,r)$ qubits.

The theorem implies a way of performing VQE given bounds on $\delta(t,r)$ using the resources of a NISQ device more efficiently. This is because implementing the smaller, effective circuit, requires fewer qubits and gates. Consider the objective of calculating the ground state energy of a two local Hamiltonian $H$. The Hamiltonian is two local in the sense that it only acts on nearest neighbors in $G_T$. Moreover, suppose that each local term $H_i$ satisfies $||H_i||_\infty \leq 1$. It suffices to estimate all $H_i$ individually to obtain an estimate of the global energy of the state by adding up the energy terms. Now suppose that we can implement each 2-qubit gate with an error $\epsilon_U$ in operator norm and can prepare each initial qubit up to an error $\epsilon_P$. This implies that the total error of implementing the causal cone and measuring each $H_i$ is bounded by $\epsilon_U N_U(t,2) + \epsilon_P N_Q(t,2)$. Thus, by only implementing the circuit from iteration $t$ to $T$, it is possible to estimate the energy of each term with an error of

$$2\delta(t,r) + \epsilon_U N_U(t,2) + \epsilon_P N_Q(t,2).$$

**Lemma.** Let $OT$ be an observable supported in a ball of radius $r$. Then

$$|tr(\Phi^*_{[t,T]}(\rho') OT) − tr(\rho_{OT})| \leq 2\delta(t,r)||OT||_\infty.$$ (2)

where $\rho = tr_B [\Phi^*_{[0,T]}(\rho_0 \otimes \rho_B)]$, which holds for all $\rho'$.

**Proof.** By the definition of $\delta(t,r)$, we see that

$$\Phi^*_{[t,T]}(OT) = O_t = cI + \delta(t,r)A,$$ (3)

where $A$ is some observable supported on supp($O_t$) satisfying $||A||_\infty \leq ||OT||_\infty$. As $\Phi^*_{[0,t-1]}$ is a quantum channel in the Heisenberg picture, $\Phi^*_{[0,t-1]}(I) = 1$ and $||\Phi^*_{[0,t-1]}||_\infty \rightarrow \infty \leq 1$ [22]. This implies that

$$\Phi^*_{[0,t-1]} \circ \Phi^*_{[t,T]}(OT) = cI + \delta(t,r)\Phi^*_{[0,t-1]}(A),$$ (4)

by combining (3) and (4).
This shows that as long as there is a $t$ such that (5) is smaller than $\epsilon_U N_U(T, 2) + \epsilon_P N_Q(T, 2)$, which corresponds to the worst case error of implementing the whole causal cone, this procedure will lead to a more efficient and less noisy way of estimating the energy. If $N_Q(t, 2) < N_Q(T, 2)$, fewer qubits are also needed when compared to generating the full state.

The stability bound in (5) readily generalizes to observables with arbitrary radius $r$ and can be further im-

| Scheme | Error | Gates | Qubits |
|--------|-------|-------|--------|
| DMR\(A\) | $\epsilon_U \lambda^{-1} D^2 + \epsilon_P \lambda^{-1} D$ | $2T D^2$ | $2T$ |
| MPS | $\epsilon_U \lambda^{-2} D^2 + \epsilon_P \lambda^{-1} D$ | $T^2 D^2$ | $TD$ |
| RI-$d$ | $\epsilon_U \lambda^{-d-1} D^{d+1} + \epsilon_P \lambda^{-d} D^d$ | $T^{d+1} D^{d+1}$ | $TD^d$ |

TABLE I. Comparison of the error estimate, number of gates and qubits necessary to estimate a two local observable on a noisy quantum computer by restricting to the causal cone, as in (2), and preparing the whole state for different interaction schemes under the assumption that $\delta(t, r) = c e^{-\lambda (T-t)}$. RI-$d$ refers to the case in which a $d$-dimensional bath interacts with a $d$-dimensional system at each iteration. For each interaction scheme, the first row corresponds to preparing the whole state. The second corresponds to only preparing the causal cone from $t_s$ to $T$, where $t_s = \lambda^{-1} \log(\epsilon_U^{-1})$. All entries of the table are only up to leading order in $D$, $T$ and $\lambda$. For $\lambda$ independent of system size, we see that it is possible to approximate all expectation values with quantum circuits of constant size.

proved to:

$$2\delta(t, r) + \sum_{k=t+1}^{T} \delta(k-1, r) \epsilon_U (N_U(k, r) - N_U(k-1, r))$$

$$+ \sum_{k=t+1}^{T} \delta(k-1, r) \epsilon_P (N_Q(k, r) - N_Q(k-1, r)).$$

In order to see this, recall that $\delta(k, r)$ measures how close the operator $O_k$ is to being proportional to the identity in the sense that there exists an operator $A_k$ with the same support as $O_k$ such that $O_k$ can be decomposed into $O_k = c \mathbb{1} + \delta(k, r) A_k$. At the $k$'th iteration, any evolution in the Heisenberg picture only acts non-trivially on $A_k$ and changes the expectation value of the observable w.r.t. to any state by at most $\delta(k, r)||O_k||$. Thus, if we actually implement a noisy version of the original evolution which is $\epsilon_U$ close to it, then we can only notice the effect of the noise in the part given by $\delta(k, r) A_k$. We conclude that each noisy unitary contributes with an error at most $\epsilon_U \delta(t, r)$, i.e. the effect of noise decreases in time if $\delta(t, r)$ decays. As there are $N_U(k, r) - N_U(k-1, r)$ new unitaries in the causal cone at the iteration, we obtain the bound. These results are intimately related with the fact that $\delta(t, r)$ and the geometry of the interactions govern the correlations present in the state produced. For $E_T$, $F_T$ two observables of disjoint support of radius $r$ and $t$ be the largest $t$ such that $E_t$ and $F_t$ have supports that intersect we can show that:

$$|\text{tr} (\rho E_T \otimes F_T) - \text{tr} (\rho E_T) \text{tr} (\rho F_T)| \leq 6\delta(t, r).$$

To demonstrate the implications of our results, we compare to the results of Refs. [9, 10] in terms of noise-
robustness and required number of gates and qubits. This comparison is summarized in Table I. We obtain a similar error bound, but are able to significantly decrease both the number of unitaries and qubits compared to the approach of Refs. [9, 10]. This is because we only require the circuit corresponding to the past causal cone to be implemented, in contrast to the circuit of the whole state. Clearly, these results also imply that it is possible to approximate these expectation values classically in some parameter regimes.

Our results provide an intuitive understanding of the stability of these computations. Each iteration contributes less to the value of expectation values, which implies that there is a small effective quantum circuit underlying the computation. Furthermore, the size of this circuit is related to the correlation length of the state and the effect of noise decreases proportionally to the amount of correlations between different regions.

In general, there are significant challenges in scaling up current qubit technologies [23–25]. The possible reduction in the number of qubits that we have shown above means that it may be possible to explore many-body quantum states with NISQ devices with substantially fewer qubits. This is of particular importance since this may bring such tasks into reach for current technology that operates with around 50 qubits [3, 4]. The possible reduction in the number of gates also reduces the necessary runtime of the circuits, which is important for hardware subject to qubit loss over time such as trapped atoms [26].

For both our approach and that of Refs. [9, 10], it is necessary to bound $\delta(t, r)$ in order to bound the error [27]. Thus, it is important to find conditions that guarantee the decay of the mixing rate and to develop protocols to estimate the mixing rate on a NISQ device.

In the translationally invariant case, one can apply the large toolbox available in the literature to estimate mixing time bounds [28–32], as further explained in the supplemental material [33]. That being said, it is important to acknowledge that obtaining rigorous mixing time bounds is a notoriously difficult problem even for classical systems [34]. But this has not kept Markov Chain Monte Carlo (MCMC) algorithms from being one of the most successful methods to simulate complex physical systems in practice [35]. Many heuristic methods are available for classical systems [36] and it would be interesting to adapt them to quantum systems. For instance, it is possible to explore that the expectation value of the observable is approximately independent of the initial state if the mixing rate is small. This can be used to devise protocols to check if the mixing rate is small [33].

We would like to thank Albert H. Werner and Carlos E. González-Guillén for helpful discussions. We acknowledge financial support from the VILLUM FONDEN via the QMATH Centre of Excellence (Grant no. 10059).

*dsfranca@math.ku.dk

[1] I. M. Georgescu, S. Ashhab, and F. Nori, Rev. Mod. Phys. 86, 153 (2014).
[2] C. Neill, P. Roushan, K. Kechedzhi, S. Boixo, S. V. Isakov, V. Smelyanskiy, A. Megrant, B. Chiaro, A. Dunsworth, K. Arya, R. Barends, B. Burkett, Y. Chen, Z. Chen, A. Fowler, B. Foxen, M. Giustina, R. Graff, E. Jeffrey, T. Huang, J. Kelly, P. Klimov, E. Lucero, J. Mutus, M. Neeley, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, H. Neven, and J. M. Martinis, Science 360, 195 (2018).
[3] J. Zhang, G. Pagano, P. W. Hess, A. Kyprianidis, P. Becker, H. Kaplan, A. V. Gorshkov, Z. X. Gong, and C. Monroe, Nature 551, 601 EP (2017).
[4] H. Bernien, S. Schwartz, A. Keesling, H. Levine, A. Omran, H. Pichler, S. Choi, A. S. Zibrov, M. Endres, M. Greiner, V. Vuletić, and M. D. Lukin, Nature 551, 579 EP (2017).
[5] J. Preskill, Quantum 2, 79 (2018).
[6] E. Farhi, J. Goldstone, and S. Gutmann, (2014), arXiv:1411.4028.
[7] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O’Brien, Nature Communications 5, 4213 (2014).
[8] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New Journal of Physics 18, 093023 (2016).
[9] I. H. Kim, (2017), arXiv:1703.00032.
[10] I. H. Kim and B. Swingle, (2017), arXiv:1711.07500.
[11] C. Schö n, E. Solano, F. Verstraete, J. I. Cirac, and M. M. Wolf, Physical Review Letters 95, 110503 (2005).
[12] M. C. Bä nzls, D. Pérez-García, M. M. Wolf, F. Verstraete, and J. I. Cirac, Physical Review A 77, 052306 (2008).
[13] D. Gross, J. Eisert, N. Schuch, and D. Perez-Garcia, Physical Review A 76, 052315 (2007).
[14] M. Fannes, B. Nachtergaele, and R. F. Werner, Communications in Mathematical Physics 144, 443 (1992).
[15] V. Giovannetti, S. Montangero, and R. Fazio, Physical Review Letters 101, 180503 (2008).
[16] G. Evenbly and G. Vidal, Physical Review B 79, 144108 (2009).
[17] O. Shehab, I. H. Kim, N. H. Nguyen, K. Landsman, C. H. Alderete, D. Zhu, C. Monroe, and N. M. Linke, (2019), arXiv:1906.00476, arXiv:1906.00476.
[18] It is straightforward to adapt our results to the case in which the the unitary channel $U_t$ depends on a classical random variable. That is, we apply some quantum channel of the form $\mathcal{T}$ which is a convex combination of unitaries respecting the scheme. This leads to a richer variety of evolutions that can be implemented (see e.g. [37]).
[19] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, (2006), arXiv:0608197.
[20] C. Schö n, K. Hammerer, M. M. Wolf, J. I. Cirac, and E. Solano, Physical Review A 75, 032311 (2007).
[21] D. Kretschmann and R. F. Werner, Physical Review A 72, 062323 (2005).
[22] D. Pérez-García, M. M. Wolf, D. Petz, and M. B. Ruskai, Journal of Mathematical Physics 47, 083506 (2006).
[23] C. Monroe and J. Kim, Science 339, 1164 (2013).
[24] B. Lekitsch, S. Weidt, A. G. Fowler, K. Mølmer, S. J. Devitt, C. Wunderlich, and W. K. Hensinger, Science
Advances 3 (2017), 10.1126/sciadv.1601540.

[25] M. Kjaergaard, M. E. Schwartz, J. Braumueller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, arXiv:1905.13641 (2019).

[26] M. Saffman, Journal of Physics B: Atomic, Molecular and Optical Physics 49, 202001 (2016).

[27] It should be noted that our approach also requires such a bound to ensure that the energy inferred from the smaller patches corresponds to a physical quantum state. This is not the case if the whole circuit preparing the many-body state is implemented.

[28] D. Burgarth, G. Chiribella, V. Giovannetti, P. Perinotti, and K. Yuasa, New Journal of Physics 15, 073045 (2013).

[29] K. Temme, M. J. Kastoryano, M. B. Ruskai, M. M. Wolf, and F. Verstraete, Journal of Mathematical Physics 51, 122201 (2010).

[30] D. Reeb, M. J. Kastoryano, and M. M. Wolf, Journal of Mathematical Physics 52, 082201 (2011).

[31] I. Bardet, (2017), arXiv:1710.01039.

[32] A. Müller-Hermes and D. S. Franca, Quantum 2, 55 (2018).

[33] See supplemental material for details.

[34] D. Levin and Y. Peres, Markov Chains and Mixing Times (American Mathematical Society, 2017).

[35] K. Binder and D. W. Heermann, in Monte Carlo Simulation in Statistical Physics (Springer International Publishing, 2019) pp. 7-70.

[36] M. K. Cowles and B. P. Carlin, Journal of the American Statistical Association 91, 883 (1996).

[37] R. Iten, R. Colbeck, I. Kukuljan, J. Home, and M. Christandl, Physical Review A 93, 032318 (2016).
Supplemental Material

This is the supplemental material to the article *Noise-robust exploration of quantum matter on near-term quantum devices*. We first discuss in more detail the growth of the causal cone, the number of unitaries, and error estimates for the examples considered in the article (Sec. I). We then review the connection between mixing times of quantum channels and the decay of the mixing rate function (Sec. II). Here, we also show that the mixing rate and the geometry of the interaction scheme bound the correlation length of sequentially generated states. Finally, we elaborate on the comparison to the results of Refs. [1, 2] (Sec. III) and describe a protocol to certify that a circuit is mixing for a given observable (Sec. IV).

I. CAUSAL CONE OF DMERA AND SEQUENTIALLY GENERATED STATES

In this section, we review the constructions for the examples considered in the main article, analyse the growth of the past causal cone and the corresponding implications for the scaling of the error of noisy implementations.

A. DMERA

Let us start by briefly recalling for the reader’s convenience the construction of DMERA states given in Ref. [2], which are depicted in Fig. 1. We start with a system consisting of one qubit. Then, at iteration \( t \) we add \( 2^t - 1 \) new qubits to the system, placing one qubit to the right of each existing qubit. Furthermore, at each iteration, we apply \( D \) layers of two-qubit unitary gates between neighboring qubits. The resulting state has a final number of \( 2^T \) qubits and it is necessary to implement \((D - 1) (2^{T+1} - 1)\) two-qubit gates to prepare the whole state.

While we add \( 2^t - 1 \) qubits in the Schrödinger picture, when looking at the Heisenberg picture of the evolution we will discard half of the qubits at each iteration. This ensures that the dynamics in the Heisenberg picture will typically be locally mixing. However, as it is the case for usual MERA, local observables have by design a causal cone that is of polynomial size in \( t \), which is crucial to all estimates in the main article. We will now discuss their growth in more detail.

![FIG. 1. Depiction of the DMERA for iterations 0 to 4. The circles (green, filled) denote the system qubits. The thick, black lines indicate where a qubit goes from one iteration to the next and the thin, gray lines indicate which qubits are neighbors at a given iteration. The digits, always next to the first qubit, indicate the iteration.](image-url)

Let us start with the number of unitaries in the past causal cone in DMERA. Recall that when looking at what happens at each iteration in the Heisenberg picture, after discarding every second qubit present in the previous iteration, we apply a unitary circuit of \( D \) layers, always with the restriction that we can only apply unitaries between qubits that are neighbors on the line. When we apply the first layer, only unitaries which act on at least one qubit in the support have a nontrivial effect. Let \( R(O_t) \) be the radius of the observable before we apply the first layer of unitaries. Then there at most \( 2R(O_t) - 1 \) nontrivial unitaries acting on the qubits in the support and two unitaries, one to the left of the support and one to the right, that act on the qubit in the left corner of the support and the first
qubit to the left of the support and analogously to the right. Thus, we conclude that as we apply the first layer, we have $2R(O_t) + 1$ unitaries acting nontrivially and the support will increase to one qubit to the right and one qubit to the left. The next layer of the unitary circuit will then act on an observable of support with radius at most $R(O_t) + 1$. Applying the same reasoning as before, we see that the total number of unitaries that act nontrivially is $2R(O_t) + 3$. We conclude that the total number of unitaries that acts nontrivially after repeating this process $D$ times is bounded by:

$$
\sum_{k=0}^{D-1} (2R(O_t) + 2k + 1) = D(2R(O_t) + D).
$$

Let us now estimate the size of the radius at each iteration to obtain a more concrete bound on the number of unitaries.

As we observed above, if at the beginning of an iteration the radius is $R(O_t)$, it will increase by $D$ and then be halved after we discard the qubits. Thus, it will go from $R(O_t)$ to at most $\left[(R(O_t) + D)/2\right] \leq (R(O_t) + D)/2 + 1$. Applying this recursive relation, we see that if the initial radius is $R(O_T)$ at iteration $t$, the radius is bounded by

$$
R(O_t) \leq R(O_T)2^{-(T-t)} + \sum_{k=t}^{T} \frac{D+2}{2^{T-k}} = R(O_T)2^{-(T-t)} + (D+2)(2-2^{-(T-t)}).
$$

Note that this implies that the radius of an observable is bounded by a constant independent of $t$. Combining the bound above on the radius of the observable with Eq. (1), we obtain that the number of unitaries added to the cone at iteration $t$ is bounded by:

$$
D(2R(O_T)2^{-(T-t)} + 2(D+2)(2-2^{-(T-t)}) + D).
$$

From this we can easily bound the total number of unitaries in the past causal cone from iteration $t$ to $T$ by summing the contribution at each step:

$$
N_U(t, R(O_T)) \leq \sum_{k=t}^{T} D(2R(O_k) + D) \leq \sum_{k=t}^{T} D(2R(O_T)2^{-(T-k)} + 2(D+2)(2-2^{-(T-k)}) + D)
$$

$$
\leq (T-t)D(2R(O_T) + 5D + 8).
$$

Let us now estimate the number of qubits in the past causal cone. At every iteration, we grow the support by at most $D$ new qubits to the left and $D$ to the right, and we start with at most $2R(O_T)$ qubits. This leads to the bound

$$
N_Q(t, R(O_T)) \leq 2R(O_T) + 2D(T-t)
$$

We will now estimate the error of implementing the past causal cone from iteration $t$ to $T$, which, as explained in the main text, can be bounded by:

$$
2\delta(t, r) + \sum_{k=t+1}^{T} \delta(k, r) \left[ \epsilon_U(N_U(k, r) - N_U(k+1, r)) + \epsilon_P(N_Q(k, r) - N_Q(k+1, r)) \right],
$$

where we assume that each unitary is implemented with an error of $\epsilon_U$ in the $1 \rightarrow 1$ norm [3] and we can initialize each qubit up to an error $\epsilon_P$, in the sense that we can prepare a state that is $\epsilon_P$ close in trace distance to the ideal one.

Let us start by estimating the error stemming from the noisy unitaries. Note that the term $N_U(k, r) - N_U(k+1, r)$ is nothing but the newly added unitaries at iteration $k$, which we bounded in Eq. (2). It follows that the contribution to the error from the noisy unitaries from iteration $t$ to $T$ is bounded by

$$
\epsilon_U \sum_{k=t}^{T} [N_U(k, R(O_T)) - N_U(k+1, R(O_T))] \delta(k, R(O_T)) \leq \epsilon_U \sum_{k=t}^{T} D(2R(O_T) + 5D + 8)\delta(k, R(O_T)).
$$

To illustrate the bound, we assume that $\delta(k, r) = e^{-\lambda(T-k)}$. Consequently,

$$
\epsilon_U \sum_{k=t}^{T} D(2R(O_T) + 5D + 8)e^{-\lambda(T-k)} = \epsilon_U \frac{D(e^{\lambda} - e^{-(T-t)\lambda})(2R(O_T) + 5D + 8)}{e^{\lambda} - 1}.
$$
One can do a similar computation for state preparation errors. As discussed above, at most 2D qubits are added to the causal cone for each iteration. Thus, the error caused by initialization between iterations \( t \) and \( T \) is bounded by:

\[
\epsilon_p \left( 2R(O_T) + 2D \sum_{k=t}^{T} e^{-\lambda(T-t)} \right) = \epsilon_p \left( 2R(O_T) + 2D \frac{(e^\lambda - e^{-(T-t)\lambda})}{e^\lambda - 1} \right) .
\] (8)

From combining equations (7) and (8), we can conclude that the error in estimating the expectation value of an observable by implementing the past causal cone from iteration \( t \) to \( T \) is bounded by:

\[
\epsilon_u \frac{D(e^\lambda - e^{-(T-t)\lambda})(2R(O_T) + 5D + 8)}{e^\lambda - 1} + \epsilon_p \left( 2R(O_T) + 2D \frac{(e^\lambda - e^{-(T-t)\lambda})}{e^\lambda - 1} \right) + 2e^{-\lambda(T-t)} .
\]

Let us now suppose we only implement the past causal cone from iterations \( t_{eu} = T - \lambda^{-1} \log(\epsilon_u^{-1}) \) until \( T \). The resulting error will then be at most

\[
\epsilon_u \left( \frac{D(e^\lambda - \epsilon_u)(2R(O_T) + 5D + 8)}{e^\lambda - 1} + 2 \right) + \epsilon_p \left( 2R(O_T) + 2D \frac{(e^\lambda - \epsilon_u)}{e^\lambda - 1} \right) .
\]

By approximating \((e^\lambda - 1)^{-1}\) by \(\lambda^{-1}\), we see that the error stemming from the noisy unitaries is at most order \(\epsilon_u D^2 \lambda^{-1}\). Similarly, the error from noisy initialization of qubits is at most order \(\epsilon_p \lambda^{-1} D\). Moreover, by inserting \(t_{eu}\) into Eq. (4), we obtain that the total number of qubits necessary to perform this computation is at most

\[
2R(O_T) + 2D \lambda^{-1} \log(\epsilon_u^{-1})
\]

and the number of unitaries that needs to be implemented is bounded by

\[
\lambda^{-1} \log(\epsilon_u^{-1}) D(2R(O_T) + 5D + 8),
\]

which follows from inserting \(t_{eu}\) into Eq. (3). Thus, under these assumptions it possible to compute local expectation values of fixed radius with noisy circuits whose error and size only depends on \(\lambda\) and \(D\), not \(T\).

### B. MPS and RI-\(d\)

Another important subclass of states are those that are sequentially generated. The most prominent example is matrix product states (MPS). Here, only one qubit interacts with the bath at each iteration. A simple generalization of this is where a group of qubits (arranged according to a \(d\)-dimensional graph) interacts with a bath (arranged according to a \(d'\)-dimensional graph) at each iteration, see Fig. 2 for an example with \(d = 0\) and \(d' = 2\). For this work, we also make the restriction that the interaction is given by a circuit of depth at most \(D\). Setting \(d = 0\) and \(d' = 1\), i.e. a qubit interacting with qubits on a line, recovers our version of MPS. We will also discuss the case of \(d = d'\) in more detail, which we will refer to as RI-\(d\). The of case \(d = d' = 1\) encapsulates examples like holographic computation discussed in [4].

We now discuss the growth of scaling of the errors in both our version of MPS and RI-\(d\). Unlike we did for DMERA, we will not fix the exact graph that models the interactions in the bath and between system and bath at each iteration and choose to focus on the scaling of the size of causal cones. More precisely, we will assume that there are constants \(C_V\) and \(C_E\) such that for every ball of radius \(r\) in the interaction graph there are at most \(C_E r^d\) edges and \(C_V r^d\) vertices inside the ball.

Let us now analyse the growth of causal cones. As it was the case with DMERA, if at the beginning of iteration \(t\) the radius of an observable is \(R(O_t)\), it will then grow to at most \(R(O_t) + D\). However, unlike for DMERA, for the interaction schemes considered here we do not discard qubits between different iterations. Thus, the radius at iteration \(t\) of an observable is bounded by \(R(O_T) + (T - t) D\). This allows us to conclude that the number of qubits in the past causal cone is bounded by:

\[
N_Q(t, R(O_T)) \leq C_V (R(O_T) + (T - t) D)^d
\]

for RI-\(d\) and \(C_V (R(O_T) + (T - t) D)\) for MPS. Let us now do a similar computation for the number of unitaries in the past causal cone. Supposing that the radius of the observable is \(R(O_t)\) at the beginning of the iteration, there are at
most $C_E(R(O_t) + 1)^d$ unitaries that act nontrivially on the first layer and the radius will grow by one. For the second layer, there will be at most $C_E(R(O_t) + 2)^d$ and the radius will again grow by one. We conclude that applying the $D$ layers will require a total of at most

$$C_E \sum_{k=0}^{D-1} (R(O_t) + k + 1)^d$$

(9)

units for iteration $t$. As $(R(O_t) + k + 1)^d$ is monotone increasing in $k$, we have that the number of unitaries added at each iteration is bounded by:

$$C_E \sum_{k=0}^{D-1} (R(O_t) + k + 1)^d \leq C_E \int_1^D (R(O_t) + x + 1)^d \, dx = \frac{C_E}{d+1} \left[ (R(O_t) + D + 1)^{d+1} - (R(O_t) + 2)^{d+1} \right]$$

(10)

where for the last inequality we used our estimate for the radius of the observable at iteration $t$ and the fact that the function $x \mapsto (x + D + 1)^{d+1} - (x + 2)^{d+1}$ is monotone increasing for $x \geq 0$ and $D \geq 1$, as can be seen by a direct inspection of its derivative. Thus, we can bound the maximum number of unitaries in the causal cone between iteration $t$ and $T$ by:

$$N_U(t, r) \leq \frac{C_E}{d+1} \sum_{k=t}^{T} \left[ (R(O_T) + (T - k + 1)D + 1)^{d+1} - (R(O_T) + (T - k)D + 1)^{d+1} \right].$$

Let us now estimate this sum. To this end, define the function $f(x) = (R(O_T) + xD + 1)^{d+1}$. By the mean value theorem there exists $\xi_k \in [T - k, T - k + 1]$ such that:

$$f(T - k + 1) - f(T - k) = f'(\xi_k) = (d + 1)D (R(O_T) + \xi_k D + 1)^d \leq (d + 1)D (R(O_T) + (T - k + 1)D + 1)^d.$$

(11)

Thus, inserting this bound into (10) it follows that

$$N_U(t, r) \leq C_E (d + 1) \sum_{k=t}^{T} D (R(O_T) + (T - k + 1)D + 1)^d \leq (d + 1)C_E D \int_1^{(T-t)+2} (R(O_T) + xD + 1)^d \, dx$$

$$= C_E \left[ (R(O_T) + (T - t + 2)D + 1)^{d+1} - (R(O_T) + D + 1)^{d+1} \right].$$

In particular, for MPS this gives a bound of

$$N_U(t, r) \leq C_E \left[ (R(O_T) + (T - t + 2)D + 1)^2 - (R(O_T) + D + 1)^2 \right].$$
We now assume that $\delta(t,r) = e^{-(T-t)\lambda}$ to bound the estimation error from implementing the past causal cone, as we did with DMERA. Recall that we bounded the number of new unitaries in the past causal cone at each iteration in Eq. (11). Once again, combining these estimates with our assumption on the mixing rate function and (5) yields a bound on the error stemming from the unitaries of at most

$$C \sum_{k=t}^{T} D(d+1)(R(O_T) + (T-k+1)D+1)^{d} e^{-\lambda(T-k)}.$$ 

Let us now estimate this sum. First, define the function

$$g(x) = (R(O_T) + xD + 1)^{d} e^{-\lambda x}.$$ 

We have:

$$g'(x) = (R(O_T) + xD + 1)^{d-1} e^{-\lambda x} (dD - \lambda (R(O_T) + xD + 1))$$

For $x \geq 0$, we see that the function is monotone increasing for

$$x \leq x_0 := \frac{1}{D} \left( \frac{dD}{\lambda} - R(O_T) - 1 \right)$$

and monotone decreasing for $x \geq x_0$. This allows us to conclude that:

$$\epsilon U C \sum_{k=t}^{T} D(d+1)(R(O_T) + (T-k+1)D+1)^{d} e^{-\lambda(T-k)} \leq \epsilon U C D(d+1) \left[ g(x_0) + \int_{x_0}^{T-t+1} g(x) dx \right] \leq 2\epsilon U C D(d+1) \int_{0}^{T-t+1} g(x) dx.$$ 

It now remains to estimate this integral. It is easy to compute the integral above using integration by parts for $d = 1$ and $d = 2$ here. For $d = 1$ we have:

$$\int_{0}^{T-t+1} (R(O_T) + (x + 1)D)e^{-\lambda x} dx = \frac{1 - e^{-\lambda(T-t+1)}}{\lambda^2} (\lambda R(O_T) + D\lambda + D) - \lambda^{-2} e^{-\lambda(T-t+1)} (D\lambda(T-t+1)), \quad (14)$$

and for $d = 2$ we obtain:

$$\frac{1 - e^{-\lambda(T-t+1)}}{\lambda^3} \left( \lambda^2 (D + R(O_T))^2 + 2\lambda D(D + R(O_T)) + 2D^2 \right) - \lambda^{-3} e^{-\lambda(T-t+1)} (2\lambda^2 D^2 (R(O_T) + D)(T-t+1) + D^2 \lambda^2 (T-t+1)^2).$$

It is then possible to obtain explicit bounds by combining the equations above with Eq. (12). But it is easy to see by direct inspection that, assuming $R(O_T) \leq D$, the error will converge exponentially fast in $(T-t)$ to $O \left( \epsilon U C \left( D\lambda^{-1} \right)^{d+1} \right)$, which is again independent of $T$. It is also possible to obtain more explicit bounds on the asymptotic behaviour of the error, i.e. with $T \to \infty$. To this end, note that $g(x) \geq 0$, thus:

$$\int_{0}^{T-t+1} g(x) dx \leq \int_{0}^{+\infty} (R(O_T) + xD + 1)^{d} e^{-\lambda x} dx = D^{-1} \int_{R(O_T)+1}^{+\infty} y^{d} e^{-\frac{\lambda}{D}(y-R(O_T))} dy$$

$$\leq D^{-1} \int_{0}^{+\infty} y^{d} e^{-\frac{\lambda}{D}(y-R(O_T))} dy = \frac{\lambda}{D} e^{-\frac{\lambda}{D}(R(O_T)+1)} \int_{0}^{+\infty} y^{d} e^{-\frac{\lambda}{D}(y)} dy = \frac{D^{d} e^{-\frac{\lambda}{D}(R(O_T)+1)}}{\lambda^{d+1}} \int_{0}^{+\infty} z^{d} e^{-z} dz = \frac{d! D^{d+1} e^{-\frac{\lambda}{D}(R(O_T)+1)}}{\lambda^{d+1}}.$$ 

This allows us to conclude that the noise stemming from the noisy unitaries is bounded by:

$$2 C \epsilon U \frac{(d+1)! D^{d+1} e^{-\frac{\lambda}{D}(R(O_T)+1)}}{\lambda^{d+1}}.$$
Similar estimates hold for the total initialization errors ($\epsilon_P$). We see that the number of qubits added at iteration $t$ is bounded by:

$$(R(O_T) + (T - t + 1)D)^d - (R(O_T) + (T - t)D)^d \leq dD(R(O_T) + (T - t + 1)D)^{d-1},$$

again using the mean value theorem. Thus, we may estimate the initialization error by:

$$\epsilon_P C_V dD \sum_{k=t}^{T} (R(O_T) + (T - k + 1)D)^{d-1} e^{-\lambda t}.$$  \hfill (16)

The attentive reader must have already realized that the expression in (12) coincides with that of (16) up to a constant if replace $d + 1$ by $d$. Thus, we may use the same estimation techniques and conclude that the error is bounded by $O \left( \epsilon_P C_V (D\lambda^{-1})^d \right)$. Moreover, we may resort to the expressions in (14) and (15) if more refined inequalities in terms of $t$ and $R(O_T)$ are desired. Thus, the total error of implementing the causal cone from $T$ to $t$ is bounded by:

$$2C_V \epsilon_P d! D^d e^{\frac{1}{d} \lambda D (R(O_T)+1)} + 2C_E \epsilon_U (d+1)! D^{d+1} e^{\frac{1}{d+1} \lambda D (R(O_T)+1)} + 2e^{-\lambda(T-t)},$$

up to corrections that are exponentially small in $T - t$.

II. MIXING RATES OF QUANTUM CHANNELS

In this section, we clarify the connections between the mixing rate function and the mixing properties of quantum channels [5].

Definition 1 (Mixing quantum channel). A quantum channel $\Lambda : \mathcal{M}_d \rightarrow \mathcal{M}_d$ is called mixing if there is a unique state $\sigma$ such that $\Lambda(\sigma) = \sigma$ and for all states $\rho$ we have that

$$\lim_{n \rightarrow \infty} \Lambda^n(\rho) = \sigma,$$

where $\Lambda^n$ denotes the quantum channel composed with itself $n$ times.

Given a mixing quantum channel $\Lambda$, the main quantity of interest is $t_1(\epsilon)$, defined as

$$t_1(\epsilon) = \inf \{n | \sup_\rho \| \Lambda^n(\rho) - \sigma \|_1 \leq \epsilon \}.$$

For $\epsilon > 0$ this quantity measures how long it takes for the quantum channel to converge, i.e., its mixing time [5, 6]. Here $\| \cdot \|_1$ corresponds to the trace norm. It is well-known that correlations in tensor network or finitely correlated states are governed by mixing properties of the transfer operator [7, 8]. We will now show this connection for completeness of the exposition.

Note that

$$\sup_\rho \| \Lambda^n(\rho) - \sigma \|_1$$

corresponds to the $1 \rightarrow 1$ norm of the linear operator $\Lambda - \Lambda_\infty$, where $\Lambda_\infty(\rho) = tr(\rho)\sigma$. It follows from duality that:

$$\| \Lambda^n - \Lambda_\infty \|_{1 \rightarrow 1} = \| (\Lambda^n)^* - \Lambda_\infty^* \|_{\infty \rightarrow \infty}$$

and $\Lambda_\infty^*(O) = tr(\sigma O)1$. Now suppose, for simplicity, that we wish to compute the expectation value of an observable $O$ supported on one qubit in $S_T$ and our interaction scheme is that of MPS. In this case, the qubits only interact with the bath at each iteration and not each other. Moreover, let us assume that the system is translationally invariant in the sense that we assume that $U_t$ is the same for all $t$. Now note that

$$O_T = \Phi_T^*(O) = tr_{S_T \rightarrow T} (U_T^* (O \otimes 1_{S_{T-1}S_B})).$$

will be an observable supported on the bath alone. Furthermore,

$$\Phi^*_t(O_{t+1}) = tr_{S_{t+1}} (U_t^* (O_{t+1} \otimes 1_{S_{t+1}S_t})).$$
Since we have assumed the action of all \( \mathcal{U}_t \) to be the same, we may define the quantum channel \( \Lambda_B^* \) from the bath to itself as

\[
\Lambda^*_B(X) = \text{tr}_{S_t A_t}( \mathcal{U}_t^* (X \otimes 1_{S_1 \cdots S_t}) ).
\]

We then have that \( \mathcal{O}_t = (\Lambda_B^*)^{T-t}(\mathcal{O}_1) \). If \( \Lambda_B \) is mixing, which is the generic case [5], we may directly bound the mixing rate with a mixing time bound on \( \Lambda_B \). Let

\[
\mathcal{B}_r = \{ O : R(O) \leq r, \| O \|_\infty \leq 1 \}.
\]

Observe that

\[
\delta(t, r) = \sup_{O \in \mathcal{B}_r} \inf_{c \in \mathbb{R}} \| \Phi^*_t [t,T](O) - c 1 \|_\infty = \sup_{O \in \mathcal{B}_r} \inf_{c \in \mathbb{R}} \| (\Lambda_B^*)^{T-t}(O) - c 1 \|_\infty.
\]

For \( \Lambda_B \) mixing, a natural choice for the constant \( c \) is given by \( \text{tr}(O_1 \sigma) \), as in this case we have:

\[
\delta(t, r) \leq \sup_{O \in \mathcal{B}_r} \| (\Lambda_B^*)^{T-t-1}(O_1) - \text{tr}(O_1 \sigma) 1 \|_\infty \leq \| \Lambda_B^{T-t-1} - \Lambda_B, \infty \|_{1 \rightarrow 1}.
\]

We conclude that in this case, \( \delta(t, r) \) can be bounded using mixing time techniques [5, 6, 9–11]. But note that these might provide a too pessimistic bound on \( \delta(l, r) \), as they do not take into account the radius of the support \( r \).

Although we made the restrictive assumption that all \( \mathcal{U}_t \) are the same, it is straightforward to adapt the arguments above to the case where they are different. This, however, implies that the sequence of quantum channels of interest is not homogeneous in time. It is, in general, known how to estimate the convergence or even certify convergence for a non-homogeneous sequence. One important exception is when the quantum channels change adiabatically in time [12]. Moreover, the results of Refs. [13, 14] seem to indicate that we should expect an exponential decay of the mixing rate function for generic local circuits of logarithmic depth in the number of qubits, but we leave this investigation for future work.

### A. Correlation length of the produced states

Here we discuss how the mixing rate function \( \delta(t, r) \) and the geometry of the interaction scheme can be used to bound the correlations present in the state produced. We measure the correlations in the state in terms of the covariance, which we introduce below.

**Definition 2** (Covariance). Let \( E, F \) be observables with disjoint support in \( G_T \). Their covariance with respect to a state \( \rho \), \( \text{cov}_\rho(E, F) \), is defined as:

\[
\text{cov}_\rho(E, F) = \text{tr}(\rho E \otimes F) - \text{tr}(\rho E) \text{tr}(\rho F).
\]

We then have:

**Proposition** (Correlations of the state). Let \( E \) and \( F \) be observables whose support is disjoint and contained in a ball of radius \( r \) and \( \rho = \Phi_{[0,T]}( \rho_0 \otimes \rho_B ) \). Moreover, let \( t_0 \) be the largest \( t \) s.t. \( E_t \) and \( F_t \) have supports that intersect. Then

\[
| \text{cov}_\rho(E, F) | \leq 6 \delta(t_0, r) \| E \|_\infty \| F \|_\infty.
\]

**Proof.** Note that for \( t > t_0 \) the supports of \( E_t \) and \( F_t \) are disjoint by definition, that is, \( \Phi^*_t [t,T](E \otimes F) \) are still product observables. By the definition of the mixing rate, there are constants \( c_E \) and \( c_F \) such that:

\[
\Phi^*_t [t,T](E \otimes F) = (c_E 1 + \delta(t, r) E'_t) \otimes (c_F 1 + \delta(t, r) F'_t).
\]

Here \( E' \) is an observable satisfying \( \| E' \|_\infty \leq \| E \|_\infty \) and whose support is contained in the support of \( E_t \). Analogous properties apply to \( F'_t \). Moreover, note that \( c_E \leq \| E \|_\infty \). Defining

\[
\tilde{C} = \Phi^*_t [t,T]((c_E 1 \otimes \delta(t, r) F'_t) + (\delta(t, r) E'_t \otimes c_F 1) + \delta(t, r) E'_t \otimes \delta(t, r) F'_t)
\]

we have that

\[
\text{tr}(\rho E \otimes F) = \text{tr}(\rho_0 \otimes \rho_B \Phi^*_0 [0,T](E \otimes F)) = c_E c_F + \text{tr}(\tilde{C} \rho_0 \otimes \rho_B).
\]
An application of the triangle inequality yields $\|\tilde{C}\|_\infty \leq 3\delta(t, r)\|E\|_\infty\|F\|_\infty$, from which we conclude
$$|\text{tr} (\rho E \otimes F) - c_E c_F| \leq 3\delta(t, r)\|E\|_\infty\|F\|_\infty.$$ (17)

A similar computation yields that
$$\Phi_{t, T}^\ast (E \otimes 1) = (c_E + \delta(t, r)E'_t) \otimes 1, \quad \Phi_{t, T}^\ast (1 \otimes F) = 1 \otimes (c_F + \delta(t, r)F'_t).$$

We, therefore, have that
$$\text{tr} (\rho E) = c_E + \text{tr} \left( \rho_0 \otimes \rho_B \tilde{C}_E \right), \quad \text{tr} (\rho F) = c_F + \text{tr} \left( \rho_0 \otimes \rho_B \tilde{C}_F \right),$$ (18)
where $\tilde{C}_E = \delta(t, r)\Phi_{t, T}^\ast (E'_t)$ and $\tilde{C}_F$ is defined analogously. From (18) we conclude that:
$$|\text{tr} (\rho E) \text{tr} (\rho F) - c_E c_F| \leq 3\delta(t, r)\|E\|_\infty\|F\|_\infty.$$

Combining the last inequality with (17) we finally have that:
$$|\text{tr} (\rho E) \text{tr} (\rho F) - \text{tr} (\rho E \otimes F)| \leq |\text{tr} (\rho E) \text{tr} (\rho F) - c_E c_F| + |\text{tr} (\rho E \otimes F) - c_E c_F| \leq 6\delta(t, r)\|E\|_\infty\|F\|_\infty.$$

\[\square\]

III. CONNECTION TO THE RESULTS OF KIM ET AL

First, we briefly review our assumptions on the noise in the implementation, which are closely related to that of Kim et al. [1, 2]. Like them, we assume that noisy versions $N_U$ of the required two qubit gates $U$ are implemented, which satisfy:
$$\|U - N_U\|_\diamond \leq \epsilon_U.$$ (19)

and the noise acts on the same qubits as $U$. Here $U$ is just the quantum channel that corresponds to conjugation with $U$ and $\|\cdot\|_\diamond$ is the diamond norm. Recall that the diamond norm is defined as
$$\|\Lambda\|_\diamond = \sup_{X \in \mathcal{M}_n \otimes \mathcal{M}_n} \frac{\|\Lambda \otimes \text{id} (X)\|_1}{\|X\|_1}$$
for a linear operator $\Lambda : \mathcal{M}_n \to \mathcal{M}_n$ and $\|\cdot\|_1$ the trace norm. The diamond norm is a natural way of quantifying the noise in our setting as it also allows us to estimate its effect on systems other than the one the unitary is acting on. However, it should be noted that as all unitaries considered in this work only act nontrivially on two qubits, the diamond norm can differ by at most a factor of 4 from $\|\cdot\|_{1 \to 1}$. That is,
$$\epsilon_U \leq 4\|U - N_U\|_{1 \to 1}.$$

We also assume that the initial state preparation is noisy. This can be modelled similarly by assuming further that all qubits are initialized in a state that is $\epsilon_P$ in trace distance to the ideal one. Let us now connect the mixing rate function of circuits to stability bounds of noisy implementations, which will allow us to recover [1, Theorem 2] in our language.

**Corollary** (Stability of noisy implementation). Let
$$\rho = \text{tr}_B \left[ \Phi_{[0, T]} (\rho_0 \otimes \rho_B) \right]$$
and $\tilde{\rho}$ be the quantum state obtained by replacing every two qubit unitary in $\Phi_t$ by a noisy counterpart satisfying (19) and every qubit initialized up to a preparation error of $\epsilon_P$. Moreover, let $O$ be an observable supported on a ball of radius $r$ and $\|O\|_\infty \leq 1$. Then for all $0 \leq t \leq T$:
$$|\text{tr} (O (\rho - \tilde{\rho}))| \leq \delta(t, r) + \sum_{k=t+1}^{T} \delta(k, r) \left| \epsilon_U (N_U(k, r) - N_U(k+1, r)) + \epsilon_P (N_Q(k, r) - N_Q(k+1, r)) \right|. $$ (20)
Proof. Let \( \tilde{\Phi}_t \) be the noisy counterpart of \( \Phi_t \). As in [1, Theorem 2], we now consider the decomposition
\[
\Phi^*[0,T] - \tilde{\Phi}^*[0,T] = \left( \Phi^*[0,t-1] - \tilde{\Phi}^*[0,t-1] \right) \circ \Phi^*[0,t] + \sum_{k=t}^{T} \tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T],
\]
with the convention that \( \Phi^*_{[t-1,0]} \), \( \Phi^*_{[T+1,T]} \) are the identity. Let us first estimate the error from the sum by estimating each summand. First, note that, as before, we have:
\[
\tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T] \circ \Phi^*[0,k-1] = \tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T] \circ \Phi^*[0,k-1] = \tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T].
\]
where once again we have \( \| A_{k+1} \|_\infty \leq \|O\|\infty \) with the same support as \( O_{k+1} \) and \( c_{k+1} \) is some constant. Moreover, \( (\Phi^*_k - \tilde{\Phi}^*_k) \) will map the identity to 0. Thus,
\[
\| \tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T] \circ \Phi^*[0,k-1] \|_\infty = \| \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T] \|_\infty.
\]
As we assumed that the noise is local, that is, it acts on the same qubits as the two-qubit gate [15] the action of \( \Phi_k \) and \( \Phi_k \) will be identical outside the support of \( A_{k+1} \). This is because both will just map the identity to the identity outside the support. This implies that only the unitary gates in the past causal cone of the observable contribute to the error and each one by each one by the support. This implies that only the unitary gates in the past causal cone contribute to the error. As there are at most \( N_U(t-1,r) \) new unitaries at iteration \( t-1 \) and at most \( N_Q(t-1,r) \) new qubits, we conclude that:
\[
\| \tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T] \|_\infty \leq \epsilon_U (N_U(k+1,r) - N_U(k,r)) + \epsilon_P (N_Q(k+1,r) - N_Q(k,r)) \tag{22}
\]
Thus, combining (21) and (22) yields:
\[
\| \sum_{k=t}^{T} \tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T] \|_\infty \leq \sum_{k=t}^{T} \delta(k+1,r) \| \tilde{\Phi}^*[0,k-1] \circ \left( \Phi^*_k - \tilde{\Phi}^*_k \right) \circ \Phi^*[k+1,T] \|_\infty \leq \sum_{k=t}^{T} \delta(k+1,r) [\epsilon_U (N_U(k,r) - N_U(k+1,r)) + \epsilon_P (N_Q(k,r) - N_Q(k+1,r))].
\]
Now, by the definition of the mixing rate function there exists an observable \( A \) such that
\[
\Phi^*[k,T](O) = cI + \delta(k,r)A
\]
with \( \|A\|_\infty \leq 1 \). Thus, we see that
\[
\left( \Phi^*[0,k-1] - \tilde{\Phi}^*[0,k-1] \right) \circ \Phi^*[k,T](O) = \delta(k,r) \left( \Phi^*[0,k-1] - \tilde{\Phi}^*[0,k-1] \right) \circ \Phi^*[k,T](A);
\]
as the identity is in the kernel of \( \Phi^*[0,k-1] - \tilde{\Phi}^*[0,k-1] \). We conclude that
\[
\| \Phi^*[0,T] - \tilde{\Phi}^*[0,T] \|_\infty \leq \delta(t,r) + \sum_{k=t}^{T} \delta(k,r) [\epsilon_U (N_U(k,r) - N_U(k+1,r)) + \epsilon_P (N_Q(k,r) - N_Q(k+1,r))],
\]
from which the claim follows. \( \square \)

The stability results of Refs. [1, 2] are captured by this corollary. For instance, the main result of Ref. [1] follows from assuming that there exist constants \( r_0, c, k, \alpha, \Delta \geq 0 \) independent of system size such that for all \( r \leq r_0 \):
\[
\delta(t,r) = c r^\alpha e^{-\gamma(T-t)} + \Delta.
\]
Optimizing
\[
t_0 = T - \frac{1}{\gamma} \log \left( \frac{\epsilon}{D r^\alpha e} \right)^2
\]
suffices to guarantee an estimate up to $O \left( D^2 \epsilon \log(\epsilon^{-1})^2 + \Delta \right)$, as in [1]. By comparing Corollary III with our main theorem (see article), we see that this stability comes from the fact that the assumptions on $\delta(t, r)$ imply that there is an "effective" circuit of constant size underlying the computation. Moreover, each iteration of the evolution can only change the expectation value by an amount that decreases with time.

This is well illustrated when we compare the bound in Eq. (20) and the one we obtained with our main result, reproduced in the supplemental material in Eq. (5). Note that the two bounds only differ by a factor of $\delta(t, r)$. This difference has a clear interpretation in light of the discussion above: in our result we allowed for an arbitrary initial state $\rho$ when implementing the past causal cone from iteration $t$ to $T$, while above the state at iteration $t$ is given by $\Phi_{[0, t-1]}(\rho_B \otimes \rho_B)$ in the noiseless version. With the previous discussion in mind, we see that any change to the state produced from iteration 0 to $t$ can only change the expectation value by $\delta(t, r)$, which explains the extra $\delta(t, r)$ factor.

IV. CERTIFYING MIXING

A close look at the proof of the main theorem shows that $\delta(t, r)$ provides a worst-case estimate for how fast the expectation values stabilize. If we are only interested in estimating the expectation value of a given observable $O$, we see that

$$\inf_{c \in \mathbb{R}} \|O_t - c1\|_{\infty}$$

gives an upper bound on the error we obtain when we estimate $\text{tr}(\rho O)$ by only implementing the circuit from iteration $t$ to $T$. Thus, it is not necessary to bound the mixing rate for arbitrary observables, which is expected to be hard in general. E.g. the results of [16] show that it is QMA-hard to determine the spectral gap [6] of certain quantum channels, which is a central quantity in determining the mixing time of quantum channels. We will therefore focus on bounding the mixing for a given observable $O$. We will show that in case $\text{supp} O_t$ is small it is possible to bound $\|O_t - c1\|_{\infty}$ on a quantum computer.

As can be seen in the proof of the main theorem, if $\delta(t, r)$ is small, then the output of the circuit is essentially independent of the initial state. Thus, it should be expected that the dependence of the expectation value of an observable $O$ on the initial state gives an estimate on the mixing time. Indeed, if we draw a state $\sigma_t$ from a state two design [17] on the support of $O_t$ and define the random variable $X_t = \text{tr}(\sigma_t O_t)$, then:

$$\left(2^{n_t} \left( \mathbb{E}(X^2_t) (2^{n_t} + 1) - 2\mathbb{E}(X_t)^2 \right) \right)^{\frac{1}{2}} \geq \|O_t - \text{tr} \left( \Phi^*_{[t, T]}(O) \right) \|_{\infty}. \tag{23}$$

Here $n_t$ is the number of qubits on the support of $O_t$. As it is possible to generate a two state design using $O(n_t \log^2(n_t))$ gates [18], equation (23) gives a protocol to measure how far each local observable is from stabilizing as long as $n_t$ is small by estimating the first and second moments of $X_t$. This protocol applies to interaction schemes for which the support of observables has a bounded radius, like DMERA.

We now discuss to derive (23) and its consequences in more detail. We start by recalling the definition of a quantum state design [17]:

**Definition 3 (State design).** A distribution $\mu$ over the set of $d$ dimensional quantum states is called a $k$-state design for some $k > 0$ if

$$\int (|\psi\rangle \langle \psi|)^{\otimes k} d\mu = \int (|\psi\rangle \langle \psi|)^{\otimes k} d\mu_U,$$

where $\mu_U$ is the (normalized) uniform measure on the set of pure quantum states.

That is, these states have the same first $k$ moments as the uniform distribution on the set of pure states. Let us now compute some relevant moments of the random quantum states: Let $|\psi\rangle$ be drawn from the uniform distribution of $d$-dimensional pure quantum states and $O$ be an observable. Moreover, define the random variable $X = \text{tr}(|\psi\rangle \langle \psi| O)$. Then:

$$\mathbb{E}(X) = \frac{\text{tr}(O)}{n}, \quad \mathbb{E}(X^2) = \frac{1}{n(n+1)} \left( \text{tr}(O^2) + \text{tr}(O)^2 \right). \tag{24}$$

This can be derived by e.g. noting that $|\psi\rangle \langle \psi|$ has the same distribution as $U|0\rangle \langle 0| U^\dagger$, where $U$ is a Haar random unitary. A simple application of the Weingarten calculus for the moments of the Haar measure on the unitary group [19, 20] yields the result. We are now ready to prove equation (23), which we restate as a lemma for the reader’s convenience:
Lemma (Checking mixing). Let $O$ be an observable and $n_t$ be the number of qubits in the support of $O_t$. Moreover, let $\sigma_t$ be drawn from a state $\rho$—design on the support of $O_t$ and denote by $X_t = \text{tr}(\Phi_{[t,T]}(\sigma_t)O)$. Then

$$
\left(2^{n_t} \left( \mathbb{E}(X^2) (2^{n_t} + 1) - 2\mathbb{E}(X)^2 \right) \right)^{\frac{1}{2}} \geq \left\| \Phi_{[t,T]}(O) - \text{tr}(\Phi_{[t,T]}(O)^2) \right\|_F \frac{1}{2^{n_t}}.
$$

Proof. Note that

$$
\left\| \Phi_{[t,T]}(O) - \text{tr}(\Phi_{[t,T]}(O)^2) \right\|_F^2 = \text{tr}(\Phi_{[t,T]}(O)^2) - 2^{-n_t} \text{tr}(\Phi_{[t,T]}(O)^2).
$$

Here $\| \cdot \|_F$ is the Frobenius norm. It follows from (24) that

$$
2^{n_t} \left( \mathbb{E}(X^2) (2^{n_t} + 1) - 2\mathbb{E}(X)^2 \right) = \left\| \Phi_{[t,T]}(O) - \text{tr}(\Phi_{[t,T]}(O)^2) \right\|_F^2
$$

if we draw $\sigma$ from the uniform distribution on states. But it is clear that the expression only depends on the second and first moments of the random variable. Thus, a state $\rho$—design satisfies the same properties. The claim then follows from the fact that $\| \cdot \|_F \geq \| \cdot \|_{\infty}$. □

As a quantum state two design of $n$ qubits can be generated with a circuit consisting of $O(n \log^2(n))$ two-qubit gates [18], we see that it is possible to check if the operator is mixing as long as it support is a small constant. Otherwise, the $2^{n_t}$ factor implies that the precision and number of samples required to check mixing is infeasible.

[1] I. H. Kim, (2017), arXiv:1703.00032.
[2] I. H. Kim and B. Swingle, (2017), arXiv:1711.07500.
[3] Strictly speaking, a bound in the diamond norm is required. However, as we will discuss in more detail later, as we only consider two qubit unitaries, they are related by a factor of four.
[4] I. H. Kim, (2017), arXiv:1702.02093.
[5] D. Burgarth, G. Chiribella, V. Giovannetti, P. Perinotti, and K. Yuasa, New Journal of Physics 15, 073045 (2013).
[6] K. Temme, M. J. Kastoryano, M. B. Ruskai, M. M. Wolf, and F. Verstraete, Journal of Mathematical Physics 51, 122201 (2010).
[7] M. Fannes, B. Nachtergaele, and R. F. Werner, Communications in Mathematical Physics 144, 443 (1992).
[8] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, (2006), arXiv:0608197.
[9] D. Reeb, M. J. Kastoryano, and M. M. Wolf, Journal of Mathematical Physics 52, 082201 (2011).
[10] I. Bardet, (2017), arXiv:1710.01039.
[11] A. Müller-Hermes and D. S. Franca, Quantum 2, 55 (2018).
[12] E. P. Hanson, A. Joye, Y. Pautrat, and R. Raquépas, Communications in Mathematical Physics 349, 285 (2017).
[13] C. E. González-Guillén, M. Junge, and I. Nechita, (2018), arXiv:1811.08847.
[14] C. Lancien and D. Pérez-García, “Correlation length in random MPS and PEPS,” (2019), arXiv:1906.11682v1.
[15] It is possible to treat the case in which the noise acts in a constant neighbourhood of the qubits similarly, but we will not discuss this scenario in order not to overcomplicate the presentation.
[16] A. D. Bookatz, S. P. Jordan, Y.-K. Liu, and P. Wocjan, Physical Review A 87, 042317 (2013).
[17] A. Ambainis and J. Emerson, in Twenty-Second Annual IEEE Conference on Computational Complexity (CCC’07) (IEEE, 2007) pp. 129–140.
[18] R. Cleve, D. Leung, L. Liu, and C. Wang, Quantum Information and Computation, Vol. 16 (Rinton Press, 2016) pp. 721–756.
[19] B. Collins and P. Śniady, Communications in Mathematical Physics 264, 773 (2006).
[20] M. Fukuda, R. Koenig, and I. Nechita, “RTNI - A symbolic integrator for Haar-random tensor networks,” (2019), arXiv:1902.08539v1.