

Average-computation benchmarking for local expectation values in digital quantum devices

Flavio Baccari^{1,2,*}, Pavel Kos^{3,4,*} and Georgios Styliaris^{3,4,*}¹*Dipartimento di Fisica e Astronomia G. Galilei, Università di Padova, I-35131 Padova, Italy*²*Padua Quantum Technologies Research Center, Università degli Studi di Padova, I-35131 Padova, Italy*³*Max Planck Institute of Quantum Optics, 85748 Garching, Germany*⁴*Munich Center for Quantum Science and Technology (MCQST), 80799 Munich, Germany*

(Received 4 August 2025; accepted 18 February 2026; published 17 March 2026)

As quantum devices progress toward a quantum advantage regime, they become harder to benchmark. A particularly relevant challenge is to assess the quality of the whole computation, beyond testing the performance of each single operation. Here, we introduce a scheme for this task that combines the target computation with variants of it, which, when averaged, allow for classically solvable correlation functions. Importantly, the variants exactly preserve the circuit architecture and depth, without simplifying the gates into a classically simulable set. The method is based on replacing each gate by an ensemble of similar gates, which when averaged together form space-time channels [P. Kos and G. Styliaris, *Quantum* 7, 1020 (2023)]. We introduce explicit constructions for ensembles producing such channels, all applicable to arbitrary brickwork circuits, and provide a general recipe to find new ones through semidefinite programming. The resulting average computation retains important information about the original circuit and is able to detect noise beyond a Clifford benchmarking regime. Moreover, we provide evidence that estimating average-computation expectation values requires running only a limited number of different circuit realizations.

DOI: [10.1103/zn6p-c8w2](https://doi.org/10.1103/zn6p-c8w2)

Introduction. Current quantum platforms allow probing many-body systems beyond sizes that can be comfortably simulated classically [1,2]. At the same time, the main limitation in the near term is hardware noise and errors, which may corrupt complex computations [3–5]. In response to this limitation, near-term and midterm quantum devices rely extensively on methods to recover the ideal outcome from noisy hardware. These range from error mitigation [6–8], applied in classical postprocessing, to error correction performed during the computation [9,10]. Reliable methods to assess the quality of quantum computations play a crucial dual role in this context. First, they serve to detect the presence of noise, providing valuable feedback that guides the improvement of error correction and mitigation strategies. Second, they help gain confidence in the reliability of the output for computations beyond the classically simulable regime.

Given a quantum computation of interest, assessing the quality of its implementation falls into the broad category of benchmarking and certification [11]. The simplest approach in this context is to test the error rate of single operations.

Methods for this range from full gate tomography [12,13] to schemes that extract specific fidelity metrics, such as randomized benchmarking [14] and variations that allow to estimate correlated errors [15–17]. However, testing single operations provides little information about the overall quality of the entire computation, especially in regimes of moderate circuit depth [18]. At the opposite end of the spectrum, verification and certification schemes check that a given device behaves correctly by testing its performance on carefully crafted computations. They normally rely on additional knowledge about the desired output of the computation and leverage it to check how closely the device reproduces it [19–22].

A highly desirable complement to previous methods is a benchmarking scheme that can test the quality of the overall computation in a regime that is as close as possible to the algorithm of interest. Such a task becomes particularly challenging in today's computations, where most experiments produce data for which the expected output is not known in advance. Motivated by this challenge, recent proposals introduced randomized motion-reversal techniques that enable estimation of the average fidelity of arbitrary circuits without prior knowledge of their ideal outputs [23,24]. While average fidelity is a powerful global metric, it remains highly desirable to complement it with more task-specific benchmarks. In this Letter, we focus on a particularly natural objective for quantum simulation experiments: assessing the accuracy of local observables at the circuit output. Current strategies rely on testing performance on simplified circuits, for which properties of the output can be computed efficiently [25–27]. Two typical choices are (1) circuits of smaller size or depth, which allow for brute force simulation and (2) circuits composed of Clifford operations [28] or matchgates [29], which

*These authors contributed equally to this work.

†Contact author: flavio.baccari@unipd.it‡Contact author: pavel.kos@mf.uni-lj.si§Contact author: georgios.styliaris@mpq.mpg.de

admit efficient classical simulation of both expectation values and measurement sampling at the output [30–32]. However, altering the circuit to be executed typically affects the noise behavior, as the latter might strongly depend on the circuit architecture and the specificity of the gates. Hence, performing checks on circuits that significantly differ from the target computations has a high chance of providing wrong quality assessments, unless specific assumptions on the noise model are made [33].

Here, we propose a benchmarking scheme that avoids any such simplifications and aims to provide a more informative testing regime applicable to any digital quantum computation. The method is based on *randomizing* the desired computation while keeping intact the executed circuit's architecture, depth, and size. The key property is that, although the individual realizations are not classically simulable, calculating few-body correlation functions of the *average* computation is classically efficient. Importantly, the value of the correlation functions carries information about the target circuit and, in general, remains nontrivial. Our schemes do not require ancilla qubits or measurements, and thus each individual realization closely resembles the target circuit. Next, we outline the general idea and then propose specific examples of such average-computation benchmarking, based on suitably chosen modifications of two-qubit gates. For instance, each gate might be randomly replaced with one of the four variants of it with equal probability.

Average-computation benchmarking. We consider the following benchmarking task: Given a circuit that one is interested in running on the device, how closely does the actual implementation approximate the ideal result of some observables of interest? We seek a method that involves executing circuits sharing the same layout, size, and depth as the original circuit. Moreover, each such circuit should, in general, not admit a trivial classical simulation.

In what follows, we introduce the notion of average-computation benchmarking and argue that it satisfies all of the above requirements. We take a digital model of computation and define the output of an arbitrary quantum circuit as

$$\mathcal{U}(|\psi_0\rangle\langle\psi_0|) = U_k \cdots U_2 U_1(|\psi_0\rangle\langle\psi_0|), \quad (1)$$

where $|\psi_0\rangle$ is the initial state and $U_i(\cdot)$ are two-qubit gates [34]. Consider also a collection of observables $\mathcal{O}_{\text{test}}$ one is interested in testing at the output. Assume now that each gate is part of an ensemble $\{U_{i,\alpha}\}_\alpha$, chosen independently from the others according to a distribution $p_i(U_{i,\alpha})$ and consider the channel resulting from the average over the whole ensemble, namely, $\mathcal{E}_i(\cdot) = \sum_\alpha p_i(U_{i,\alpha}) U_{i,\alpha}(\cdot) U_{i,\alpha}^\dagger$. By fixing a sample over gate realizations, which we call a round, and gathering data from the output of the corresponding circuits, one can estimate the expectation values of interest,

$$\langle \mathcal{O}_{\text{test}} \rangle_{\mathcal{U}^{[r]}} = \text{Tr}(\mathcal{O}_{\text{test}} \mathcal{U}^{[r]}(|\psi_0\rangle\langle\psi_0|)), \quad (2)$$

where $\mathcal{U}^{[r]}$ is the circuit evolution on the r th round. This is defined by the two-qubit unitaries $U_{k,\alpha_k}^{[r]} \cdots U_{2,\alpha_2}^{[r]} U_{1,\alpha_1}^{[r]}$, where the α variables represent the fixed outcome of the random sampling for the given round and we also adopted the shorthand notation $\langle \mathcal{O}_{\text{test}} \rangle = \{\langle \mathcal{O}_1 \rangle, \langle \mathcal{O}_2 \rangle, \dots\}$. We then define the results of the average computation to be $\mathbb{E}_{\mathcal{U} \sim p}[\langle \mathcal{O}_{\text{test}} \rangle_{\mathcal{U}}]$.

The main idea behind average-computation benchmarking is that the same average can be computed in two different ways. On the quantum computer side, one can estimate the average via the mean over rounds

$$\langle \mathcal{O}_{\text{test}} \rangle_{\text{avg}}^{\mathcal{Q}} = \frac{1}{R} \sum_{r=1}^R \langle \mathcal{O}_{\text{test}} \rangle_{\mathcal{U}^{[r]}}. \quad (3)$$

On the classical side, we can exploit the linearity of the average to rewrite

$$\begin{aligned} \langle \mathcal{O}_{\text{test}} \rangle_{\text{avg}}^{\mathcal{C}} &= \text{Tr}(\mathcal{O}_{\text{test}} \mathbb{E}_{\mathcal{U} \sim p}[\mathcal{U}] (|\psi_0\rangle\langle\psi_0|)) \\ &= \text{Tr}(\mathcal{O}_{\text{test}} \mathcal{E}_{\text{avg}}(|\psi_0\rangle\langle\psi_0|)), \end{aligned} \quad (4)$$

where $\mathcal{E}_{\text{avg}}(|\psi_0\rangle\langle\psi_0|) = \mathcal{E}_k \cdots \mathcal{E}_2 \mathcal{E}_1(|\psi_0\rangle\langle\psi_0|)$ represents the average computation as a nonunitary evolution defined by the concatenation of the average channels $\mathcal{E}_i$. Note that, to form the individual channels, we used the fact that p_i and p_j are independent if $i \neq j$. We claim that, for certain testing sets $\mathcal{O}_{\text{test}}$, the averaging ensemble $p(U_{i,\alpha})$ can be chosen so that evaluating the average-computation expectation values $\langle \mathcal{O}_{\text{test}} \rangle_{\text{avg}}^{\mathcal{C}}$ can be done with an efficient classical algorithm. Remarkably, this remains true even for cases in which none of the expectation values $\langle \mathcal{O}_{\text{test}} \rangle_{\mathcal{U}^{[r]}}$ for the single circuit runs can be computed efficiently beyond evaluating them on the quantum device. In what follows, we illustrate concrete examples of ensembles that lead to an efficient classical computation of the average expectation values. For simplicity, we particularize to one-dimensional (1D) brickwall circuit layouts and testing set $\mathcal{O}_{\text{test}}$ involving only few-body correlators. However, our method can be straightforwardly adapted to higher spatial dimensions, as illustrated in the Appendix. Even these simple correlators are already believed to be difficult to compute for an arbitrary polynomial-sized quantum circuit, with only a few special exceptions [28,35–37].

Space-time averaging. As an additional requirement, we want to define benchmarking examples whose average correlations retain some information about the single instances of the ensemble. This is to be contrasted, for instance, to taking Haar-random two-qubit unitaries U_i . In that case, the average value of any correlation function vanishes, independently of the original circuit to be benchmarked. We pursue a more informative alternative by defining ensembles where the average channels, $\mathcal{E}_i$, are space-time channels [38]. These remain valid quantum channels (i.e., completely positive and trace preserving) upon switching the roles of space and time and lead to classically efficiently computable correlators.

Space-time channels subsequently play an important role in our scheme, so we recall their definition and some basic properties. For convenience, we will use vectorization, where we map operators to vectors using $|m\rangle\langle n| \xrightarrow{\text{vec}} |m\rangle \otimes |n\rangle$ [38]. For instance, the Hilbert-Schmidt normalized identity operator is mapped to a Bell pair

$$\frac{1}{\sqrt{2}} |\mathbb{1}_2\rangle = \frac{1}{\sqrt{2}} \sum_{r=0}^1 |r\rangle \otimes |r\rangle = |\odot\rangle = \mathbb{J}. \quad (5)$$

After vectorizing density matrices, a two-qubit channel (superoperator) $\mathcal{E}$ with Kraus operators K_α is mapped to an

operator

$$\mathcal{E} \xrightarrow{\text{vec}} \sum_{\alpha} K_{\alpha} \otimes K_{\alpha}^* = \text{diag}, \quad (6)$$

where we also introduced usual graphical tensor network notation [39]. Using this notation, trace preservation and unitality are denoted as

$$\text{diag} = \text{line}, \quad \text{diag} = \text{line}, \quad (7)$$

respectively. Space-time channels arise by defining the analogous conditions in the spatial direction, that is, left and right space unitality conditions

$$\text{diag} = \text{line}, \quad \text{diag} = \text{line}. \quad (8)$$

In particular, we define four-way space-time channels as those satisfying all four conditions, while three-way space-time channels those satisfying Eq. (7) and at least one of Eq. (8).

In Ref. [38], it was shown how these conditions, for properly chosen input states, lead to efficiently computable local observables and correlation functions. We present the relevant cases in the Appendix. Here, we consider a nonexhaustive list sharing the feature that the average-computation expectation value is generally nonzero and retains dependence on the input circuit. Generally, the structure of correlations is governed by the so-called transfer matrices,

$$\mathcal{M}_+ = \text{diag}, \quad \mathcal{M}_- = \text{diag}, \quad (9)$$

which are single-qubit quantum channels that arise from the corresponding average channel $\mathcal{E}$. The simplest result concerns both three- and four-way channels a test set $\mathcal{O}_{\text{test}}$ of single-site traceless operators $O^{(T)}$ acting on site T , where T is also the circuit depth. For properly chosen input states, the average computation reduces to an expression of the form $\langle O^{(T)} \rangle_{\text{avg}} = \langle O | \prod_{k=1}^T (\mathcal{M}_+)_k | \sigma_X \rangle$, where k runs over average channels in a straight-line path connecting the first-input qubit to the output qubit at location T . Thus, classically evaluating the expression for $\langle O^{(T)} \rangle_{\text{avg}}$ amounts to matrix multiplication; crucially, the dimension of each matrix $\mathcal{M}$ is only 4 (for the case of qubits considered here), independent of the system size.

Average circuits composed of three-way channels also allow for efficient classical computation of nearest-neighbor three-site observables, $\mathcal{O}_{\text{test}} = \{O^{(i,i+1,i+2)}\}_{i=0}^{L-3}$. Finally, average circuits leading to four-way channels allow for efficient classical computation of all two-body correlators at distance $2T+1$, namely, $\mathcal{O}_{\text{test}} = \{O^{(i,i+2T+1)}\}_{i=0}^{L-2T-2}$. We now show explicit choices of ensembles leading, for any unitaries in the original circuit, either to four- or three-way average channels.

Four-way averaging. Consider a general parametrization of the two-qubit unitaries in the circuit

$$U_i = (W_A \otimes W_B) \exp \left(i \sum_{\alpha=X,Y,Z} \theta_{\alpha}^{(i)} \sigma_{\alpha} \otimes \sigma_{\alpha} \right) (V_A \otimes V_B), \quad (10)$$

where $W_{A/B}$, $V_{A/B}$ are arbitrary single-qubit unitary operations [40] and σ_{α} are Pauli matrices; note that this simple form is specific to qubits. We rewrite the $\vec{\theta}^{(i)}$ parameters for the two-qubit part as $\vec{\theta}^{(i,++)} = (\frac{\pi}{4} + \delta_x, \frac{\pi}{4} + \delta_y, \theta_z)$. Define the unitary $U_i^{(\pm\pm)}$ as the one that keeps the same single-qubit part of Eq. (10) and replaces the vector $\vec{\theta}^{(i)}$ with $\vec{\theta}^{(i,\pm\pm)} = (\frac{\pi}{4} \pm \delta_x, \frac{\pi}{4} \pm \delta_y, \theta_z)$. A simple ensemble resulting in a four-way average channel for each site i is the uniform distribution over the four unitaries $\{U_i^{(++)}, U_i^{(+-)}, U_i^{(-+)}, U_i^{(--)}\}$. This distribution is, by construction, symmetric around the dual-unitary point $\delta_x = \delta_y = 0$ [37]; this property is enough to guarantee that the resulting channel (after averaging) is four-way unital [38]. Notice that the introduced ensemble always includes the original gate, corresponding to the $(++)$ choice.

An alternative ensemble leading to a four-way average channel comes from a strategy that resembles Pauli twirling [41]. Instead of replacing gates in the circuit, the ensemble involves applying single-qubit Pauli gates before and after each U_i . In particular, one can directly verify that the average channel defined as

$$\text{diag} = \frac{1}{4} \text{diag} + \frac{1}{12} \sum_{\alpha,\beta=X,Y,Z} \text{diag}, \quad (11)$$

is four-way for any choice of unitary U_i , where we defined $\text{diag} = U_i \otimes U_i^*$ and $\text{diag} = \sigma_{\alpha} \otimes \sigma_{\alpha}^*$. Alternatively, the Pauli matrices can be applied to the opposite pair of legs, or one can take an arbitrary convex combination of the two strategies.

Three-way averaging. A unitary ensemble leading to a three-way average channel can be obtained with a simple Pauli twirling procedure applied to one of the output legs of the original unitary U_i . More precisely, the average channel resulting from

$$\text{diag} = \frac{1}{4} \sum_{\alpha=\mathbb{I},X,Y,Z} \text{diag}, \quad (12)$$

is a three-way channel (right unital) for any input unitary. The averaging procedure corresponds to effectively applying a fully depolarizing channel to one of the outputs of each gate in the circuit.

Another approach is to first simplify the dynamics by adding dephasing channels after each gate, implemented by applying σ_Z with 50% probability. This simplifies the unitary channel to a reduced channel with only nine parameters [38]. To make the resulting channel three-way (four-way) unital, one (two) of these parameters needs to vanish in the convex combination of such channels. Concretely, we can select a new unitary U'_i such that the target parameter changes sign. The combined channel, formed from an equal mixture of U_i and U'_i together with dephasing, yields a three-way unital channel. A suitable U'_i can be found systematically using the equations for the aforementioned parameters, which are reported in Appendix B1 of Ref. [42].

Relation to the original circuit. Let us elaborate on how the proposed average computations retain information about the original circuit. To make a quantitative assessment, we analyze how closely the Pauli coefficients

$(\mathcal{E}_i)_{\beta_1\beta_2,\alpha_1\alpha_2} = \text{Tr}[\sigma_{\beta_1} \otimes \sigma_{\beta_2} \mathcal{E}_i(\sigma_{\alpha_1} \otimes \sigma_{\alpha_2})]$ of the average channel are related to the coefficients of each original unitary U_i .

The first relevant components to study are the set of coefficients forming the transfer matrices of Eq. (9), since they determine nonzero correlations in space-time channels (see the Appendix). Interestingly, in the averaging procedures proposed above, the coefficients of the average channel's transfer matrix are identical to U_i 's, or differ only by a rescaling factor $1/3 \leq x \leq 1$, which depends on the specific ensemble (see Sec. I of the Supplemental Material [43] for details). This perspective suggests a more systematic search for averaging strategies by identifying averaging ensembles that preserve the Pauli coefficients of the input unitary as much as possible.

In Sec. I of the Supplemental Material [43], we show how to formulate the problem of searching for a given rescaling supermap as a semidefinite program (SDP). The program allows for searching for averaging strategies that output four- or three-way channels with optimal values for the rescaling factors of interest. It can be used to show that no linear supermap can act as a projection on space-time channels such that the resulting channel satisfies Eq. (8) while keeping the remaining Pauli coefficients $(\mathcal{E}_i)_{\beta_1\beta_2,\alpha_1\alpha_2}$ the same. Notice also that not all proposed averaging strategies can be written as supermaps. For instance, the reflection ensemble $U_i^{(\pm\pm)}$ cannot, since the resulting channel does not depend linearly on the input unitary.

Once a supermap is found, it is important to find a way to test whether it can be implemented as an averaging strategy. Section II of the Supplemental Material [43] provides a linear program [46] to check whether a supermap can be realized as averaging over the application of single-qubit Pauli gates before and after the input unitary. In the same section, we show how the combination of the SDP and the linear program allows us to discover the simple averaging strategies of Eqs. (11) and (12).

Noise benchmarking. Next, we showcase that expectation values on average computations are sensitive to noise in the quantum circuit. Therefore, a discrepancy between the analytical prediction and the experimental value signals the presence of unaccounted-for noise. As an illustrious concrete example of an error that standard Clifford benchmarking would fail to detect, we consider a brickwork circuit, composed of repeating uniformly the gate

$$U_{(n)} = \text{CNOT}_{12}(u_n \otimes u_n) \text{CNOT}_{21}(u_n \otimes u_n) \text{CNOT}_{12}(u_n \otimes u_n), \quad (13a)$$

where we consider three possible options

$$u_1 = HT^2HTH, \quad u_2 = HT^2HTHTH, \quad u_3 = HT^2HTHT, \quad (13b)$$

and a concrete noise model, i.e., a consistent but incorrect rotation of the T-gate $T = \text{diag}(1, e^{i\phi})$, where $\phi = \pi/4$ corresponds to no error. Figure 1 shows the analytically computed expectation values of a local observable versus ϕ using our three-way averaging scheme of Eq. (12).

An alternative scenario that fits well with the proposed average-computation benchmarking comes from Pauli noise. Consider a circuit realization where we assume that every layer is followed by a gate-independent error channel $\mathcal{E}_{\text{noise}} =$

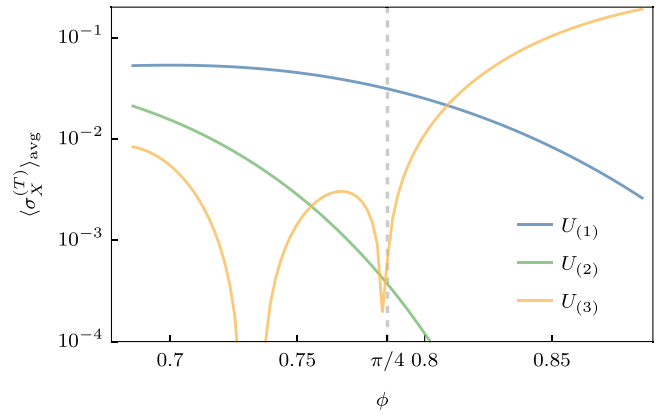


FIG. 1. Sensitivity to coherent noise. Expectation values are plotted vs the rotation angle in the imprecise T-gate $T = \text{diag}(1, e^{i\phi})$ after five layers of the circuit. Dashed line marks the correct value $\phi = \pi/4$. The three gates are given in Eq. (13). This demonstrates that our benchmarking scheme can detect this type of coherent error, in contrast to Clifford benchmarking. We take the initial state as explained in the Appendix and operator σ_X at position T .

$\sum_{\vec{\alpha}} p_{\vec{\alpha}} \sigma_{\vec{\alpha}}(\cdot) \sigma_{\vec{\alpha}}$, which is a mixture of (possibly many-qubit) Pauli errors. Such a model is widely used for error characterization, error tailoring, and error mitigation in current experiments [17,27,41,47,48]. We show in Sec. IV of the Supplemental Material [43] that, under such a noise model, any averaging strategy leading to space-time average channels allows for efficient computation of noisy expectation values as well. In this case, testing the classically computed values with the ones from the quantum device serves as a benchmark for the given noise model, providing an additional use case for average-computation benchmarking.

Practical implementation. We now discuss the sample complexity of estimating the average expectation value $\langle \mathcal{O}_{\text{test}} \rangle_{\text{avg}}$ from the quantum device. Assuming $\|\mathcal{O}_{\text{test}}\| \leq 1$, Hoeffding's inequality directly implies that M independent samples suffice to estimate the expectation up to an additive error $\epsilon \sim 1/\sqrt{M}$ —independent of the system size L . Building on this, our exact numerical simulations suggest that the variance is, at least in the tested cases, actually much smaller than one, providing evidence that only a few shots are sufficient to estimate the mean expectation value. We simulated the first procedure explained after Eq. (9) and in the Appendix, i.e., for computing $\langle \mathcal{O}^{(x)} \rangle_{\text{avg}}$. Rather than looking just at the mean value (which can be computed efficiently), we analyzed the entire distribution of the measurement outcomes for a local observable at sites $x = T$ and $x = T - 1$, where in the latter the average expectation value is zero. The circuit gates were sampled according to a four-way reflection ensemble $U_i^{(\pm\pm)}$.

In Fig. 2, we plot the standard deviation of the histograms versus circuit depth (main panel) and the distribution of the resulting expectation values for each circuit realization as a histogram (inset). The results suggest that the standard deviation of the expectation values remains small and does not grow significantly with circuit depth. Consequently, the number of samples required to estimate the average expectation value with constant precision remains small. These estimated averages can then be compared to the corresponding analytical results for benchmarking purposes.

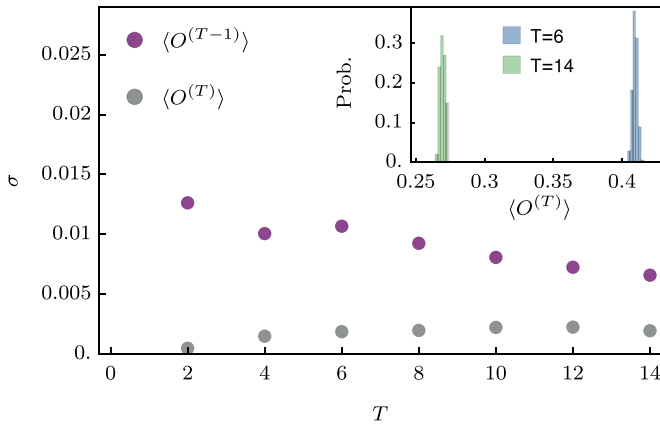


FIG. 2. Main: Standard deviation σ of a local observable's expectation value vs circuit depth (T) for operators at position $T - 1$, where the average is zero, and at position T . The standard deviation is calculated over different circuit realizations, where gates are sampled according to a four-way reflection ensemble $U_i^{(\pm\pm)}$. Inset: The full distribution of the expectation values $\langle O^{(T)} \rangle$ of different circuit realizations for depths $T = 6, 14$. The parameters of $U_i^{(\pm\pm)}$ were set to generic values not too far from dual unitarity, $\theta_z = 0.6$, $\delta_x = \delta_y = 0.05$, and we used fixed randomly chosen single-site gates and a fixed operator. The particular choice of single-site gates does not significantly alter the distribution. Their specific values are reported in Sec. III of the Supplemental Material [43].

Second, we analyze how the correlations used for benchmarking depend on the original circuit. As discussed before, correlations in our average computations are determined by suitable multiplication of the transfer matrices (9), which are inherited from the original unitaries in the circuit. By following the analysis in Ref. [38], the second leading eigenvalues of those matrices determine the speed of decay of the average correlations used for benchmarking. We could not find any example of averaging ensemble leading to a transfer matrix with two degenerate eigenvalues equal to 1, beyond the trivial case of dual-unitary gates, which would make the initial circuits already solvable. Hence, we typically expect the average correlations $\langle O_{\text{test}} \rangle_{\text{avg}}^C$ to decay exponentially with the circuit depth, likely requiring an increasing number of samples to benchmark. Averaging strategies that rescale the initial transfer matrix, such as Eq. (11), will increase the decay speed by the rescaling factor. This makes average-computation benchmarking mostly suited to low-depth quantum circuits, as those addressed in typical near-term quantum computations.

Conclusions. In this Letter, we introduced average-computation benchmarking of quantum devices. The method involves replacing each gate in the circuit with one chosen from an ensemble of similar gates. While each individual random circuit likely remains as computationally complex as the original, the expectation values of local observables, when averaged over the ensemble, become classically tractable.

Our scheme offers important advantages over existing methods. It assesses the performance of the entire computation without simplifying the original circuit or altering its size and depth. We also demonstrated that the method is sensitive to coherent errors that evade standard Clifford-based protocols

and that the number of experimental runs required to estimate the benchmark values is manageable.

Poor agreement with average-computation expectation values provides a direct signature of uncompensated noise in the device. As with other benchmarking schemes, however, passing the test does not strictly speaking guarantee a flawless computation, since certain errors may be suppressed by the averaging procedure. Our method should be thought of as part of a broader set of complementary tests that strengthen confidence in the validity of the executed computation.

Beyond hardware characterization, the benchmarking procedure can directly be used to evaluate the performances of classical simulation algorithms for many-body dynamics and quantum computation.

Generalizing our proposed approach to qudits is straightforward. While the reflection strategy is not easily generalizable due to the lack of a standard decomposition for general qudit gates, our other strategies are readily adaptable. The generalization would also provide an alternative benchmarking scheme for qubits by grouping several sites together to form effective qudits.

The local correlators used in our examples decay exponentially with circuit depth, making the scheme most suitable for near- and midterm applications. A promising future direction is to explore benchmarking schemes based on nondecaying signals, such as out-of-time-ordered correlators (OTOCs). This would require an extension of the current framework, as computing OTOCs involves “two-replica” quantities that are not solvable within the class of space-time channels presented here.

Acknowledgments. We thank Bruno Bertini and Ignacio Cirac for valuable discussions. The work at MPQ was partly funded by THEQUCO as part of the Munich Quantum Valley, which is supported by the Bavarian state government with funds from the Hightech Agenda Bayern Plus. P.K. acknowledges financial support from the Alexander von Humboldt Foundation. F.B. acknowledges financial support from the ICSC—Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union—NextGenerationEU and from the European Union’s Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Action for the Project No. 101148556 (ENCHANT). The views and opinions expressed here are solely those of the authors and do not necessarily reflect those of the funding institutions. Neither of the funding institutions can be held responsible for them.

The authors of this paper were ordered alphabetically.

Data availability. Raw data plotted in Fig. 2 are available from a Zenodo repository [49].

Appendix: Benchmarking schemes. Here, we detail how four- and three-way unital channels simplify the calculation of certain local expectation values at the output of an average computation, where each unitary in the circuit is replaced by a channel. We use the diagrammatic folded representation introduced in the *space-time averaging* section (see also Ref. [38]).

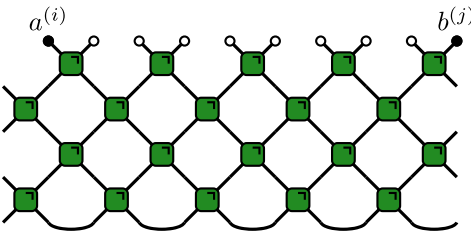
As in the main text, we formulate the quantities to be tested in terms of observables. The schemes can, however, be directly adapted to yield the corresponding reduced state by contracting the same diagrams with the observables removed.

Initial states. Our procedure requires compatible solvable initial states [38], or an averaging procedure that transforms the initial states into solvable ones. States that are compatible with our procedure are solvable matrix product density operators (MPDOs) [38]. Their defining characteristic is that they are trace-preserving maps from left to right or vice versa, where we group the left (right) physical and auxiliary indices together. In this Letter, we will limit ourselves to a simple example, a product state on sites $0, \dots, L-1$ composed of Bell pairs, $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$:

$$\rho(0) = \rho_{\text{Bell}}^{\otimes L/2}, \quad \rho_{\text{Bell}} = |\Phi^+\rangle\langle\Phi^+|. \quad (\text{A1})$$

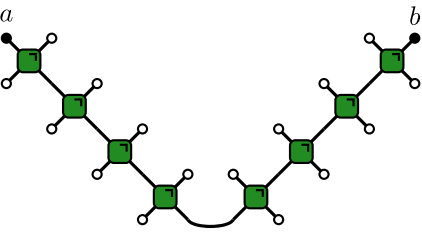
The state preparation can be treated as an additional single layer of gates, as compared with the target circuit.

Two-body correlations: Four-way channels. Once a solvable initial state ρ_0 is prepared, we can compute physical observables. First, we consider the two-body correlations between local operators $a^{(i)}$ and $b^{(j)}$ at position i and j , respectively, captured by two-body operator $O^{(i,j)} = a^{(i)}b^{(j)}$. After evolution by T layers of circuit, they are given by $\langle O^{(i,j)} \rangle_{\text{avg}} = \text{Tr}(O^{(i,j)}\rho(T))$, where we assume $j > i$. $\rho(T)$ results from a quantum evolution of T layers with a brickwall circuit composed of local quantum channels. Graphically, we express the two-point equal time correlation function as



$$\langle O^{(i,j)} \rangle_{\text{avg}} = \text{Diagram (A2)} \quad (\text{A2})$$

In the presence of left and right unitality and trace preservation, which is the case for four-way channels, the diagram simplifies considerably. In particular, for $j > i + 2T + 1$ and our specific solvable initial states, the correlations are equal to zero. Instead, for $j = i + 2T + 1$ the diagram can be simplified to [38]



$$\langle O^{(i,i+2T+1)} \rangle_{\text{avg}} = \text{Diagram (A3)} \quad (\text{A3})$$

First, we simplified from the top using trace preservation, which created backward light cones from the local operators. Then, we use left and right unitality to simplify the diagram from the sides. For $j < i + 2T + 1$, the diagram is almost the same, except that the initial Bell pair is changed to $(\text{vec } \mathcal{E})|\bigcirc\bigcirc\rangle$. In the presence of unitality, which is a property of all our average channels, this also simplifies to zero.

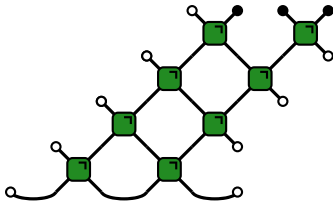
Equation (A3) is an effective 1D diagram that can be contracted efficiently. The final expression can be written using

the transfer matrices from Eq. (9) as

$$\langle O^{(i,i+2T+1)} \rangle_{\text{avg}} = \langle ab | \prod_{k=1}^T (\mathcal{M}_-)_k \otimes (\mathcal{M}_+)_k | \rho_{\text{Bell}} \rangle, \quad (\text{A4})$$

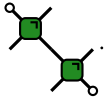
where we added the subscript k to denote that $\mathcal{M}_{\pm}$ may be different in different layers.

Nearest-neighbor k -body correlations: Three-way channels. The simplification from the previous section cannot be achieved without both left and right unitality. Nevertheless, the case of three-way channels is still interesting and tractable. Let us assume the channel is unital, trace preserving, and right unital. In this scenario, we can consider the expectation value of an observable acting on three consecutive sites: $\langle O^{(i,i+1,i+2)} \rangle_{\text{avg}} = \text{Tr}(O^{(i,i+1,i+2)}\rho(T))$. For this calculation, we simplify the corresponding diagram from the top (using trace preservation) and from the right (using right unitality), leading to the following diagram:



$$\langle O^{(i,i+1,i+2)} \rangle_{\text{avg}} = \text{Diagram (A5)} \quad (\text{A5})$$

where the black dots denote the position of the observable. This effectively one-dimensional diagram can be efficiently contracted using repeated actions of two-qubit quantum channels



$$\mathcal{M}_+^2 = \text{Diagram (A6)} \quad (\text{A6})$$

The final expression reads

$$\langle O^{(i,i+1,i+2)} \rangle_{\text{avg}} = \langle O | (\mathbb{1} \otimes \mathcal{E}_L) \prod_{k=1}^{T-1} (\mathcal{M}_+^2)_k (\mathcal{M}_R \otimes \mathbb{1}) | \rho_{\text{Bell}} \rangle, \quad (\text{A7})$$

where O is a three-site operator, and

$$\mathcal{M}_R = \langle \bigcirc |_L (\text{vec } \mathcal{E}) | \bigcirc \rangle_L, \quad \mathcal{E}_L = (\text{vec } \mathcal{E})_T | \bigcirc \rangle_R. \quad (\text{A8})$$

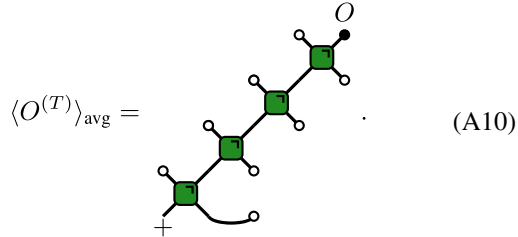
$(\text{vec } \mathcal{E})_i$ denotes the vectorized local channel in layer i and L, R denote the left or right site. Notice that a similar procedure can be used to simplify the calculation of any nearest-neighbor k -body correlations $\langle O^{(i,\dots,i+k-1)} \rangle_{\text{avg}} = \text{Tr}(O^{(i,\dots,i+k-1)}\rho(T))$. The generalized expression will involve transfer matrices acting on more and more qubits, hence resulting in a contraction effort that scales exponentially with k . Hence, the computation of correlations will remain efficient, as long as one considers a finite k .

Single-body expectation values: Three-way channels and nontranslational invariant initial state. An alternative approach is to evaluate quantities that mimic infinite temperature correlation functions. This approach, proposed in Ref. [50], works for both four-way and the aforementioned three-way unital channels. It uses an initial state of plus state $|+\rangle = 1/\sqrt{2}(|0\rangle + |1\rangle)$ and Bell pairs under open boundary

conditions:

$$\rho(0) = |\Psi_0\rangle\langle\Psi_0|, \quad |\Psi_0\rangle = |+\Phi^+\Phi^+\dots\rangle. \quad (\text{A9})$$

We then consider the expectation value of a single-site observable at site T (first site is at position 0) [51], namely, $\langle O^{(T)} \rangle_{\text{avg}} = \text{Tr}(O^{(T)}\rho(T))$. The corresponding diagram can be simplified from the top (using trace preservation) and from the right (using right unitarity), yielding



$$\langle O^{(T)} \rangle_{\text{avg}} = \text{Diagram} \quad (\text{A10})$$

This expression matches the infinite temperature space-time two-point correlation function $\text{Tr}(O^{(T)}(\sigma_X)^{(0)}(T)) = \langle O | \prod_{k=1}^T (\mathcal{M}_+)_k | \sigma_X \rangle$ [37,50]. In contrast to $2T$ single-site transfer matrices in Eq. (A3), it uses only T of them. As a result, the correlations decay to a similar level at twice the circuit depth. This slower decay is beneficial for benchmarking purposes, as the signal persists longer and is easier to measure in deeper circuits.

More general initial states. More general initial states can also be accommodated. This can be achieved in two ways: (1) by taking convex combinations of different initial states to form solvable states, or (2) by applying a layer of local channels that transforms the initial state into a solvable MPDO. A trivial example of the latter is a layer of totally depolarizing channels.

As an illustrative example of method (1), we consider an initial pure state described by an injective matrix-product states (MPS) with two-dimensional physical and bond dimensions. Here, we can think of the MPS tensor as a unitary gate on two qubits. Then, we can use the reflection procedure developed for unitary gates (four-way averaging in the main

text) to construct a solvable MPDO. When the dimensions of the auxiliary and physical spaces are different, finding these “reflections” is more difficult, but we expect that a modification of the SDP procedure outlined in Sec. I of the Supplemental Material [43] for this task is possible.

Benchmarking schemes in 2 + 1D. These ideas can be extended to two-dimensional systems. This requires generalizing dual unitarity to two dimensions, known as ternary unitarity [52–54], and combining it with the concept of space channels.

In two dimensions, we first focus on brickwall evolution, which is composed of plaquette gates acting on four sites arranged in two layers [52]. These gates are ternary unitary if, in addition to normal unitarity, they remain unitary when either of the two spatial dimensions is reinterpreted as the time direction.

Analogous to the 1D case, we can generalize these ideas to local quantum channels and construct plaquette channels that satisfy up to six unitarity conditions. These conditions are sufficient to simplify the correlation functions, ensuring they are nonzero only along the light rays, exactly as for the unitary case in Ref. [52].

A randomizing procedure leading to average six-way unital channels can be easily applied to a concrete circuit architecture on a square lattice. The circuit is composed of successive brickwall layers of the form

$$\mathcal{U} = U_{oo,\text{vert}} U_{oo,\text{horz}} U_{ee,\text{vert}} U_{ee,\text{horz}}, \quad (\text{A11})$$

where $U_{ee,\text{horz}}$ ($U_{oo,\text{horz}}$) are tensor products of arbitrary two-qubit unitaries acting on sites with even (odd) valued x coordinate and their right neighbors, and $U_{ee,\text{vert}}$ ($U_{oo,\text{vert}}$) between even (odd) valued y coordinate with their upper neighbors. Then, applying the reflection strategy to each of the gates results in four-way unital two-qubit channels. Combining vertical and horizontal interactions of even (odd) layer into plaquette channels, such as local channels arising from $U_{oo} = U_{oo,\text{vert}} U_{oo,\text{horz}}$, results in six-way unital channels.

We believe that more elaborate schemes are possible, but they are beyond the scope of this work.

-
- [1] A. Morvan, B. Villalonga, X. Mi, S. Mandra, A. Bengtsson, P. Klimov, Z. Chen, S. Hong, C. Erickson, I. Drozdov, *et al.*, Phase transitions in random circuit sampling, *Nature (London)* **634**, 328 (2024).
 - [2] A. D. King, A. Nocera, M. M. Rams, J. Dziarmaga, R. Wiersema, W. Bernoudy, J. Raymond, N. Kaushal, N. Heinsdorf, R. Harris, *et al.*, Computational supremacy in quantum simulation, *Science* **388**, 199 (2025).
 - [3] F. Pan, K. Chen, and P. Zhang, Solving the sampling problem of the sycamore quantum circuits, *Phys. Rev. Lett.* **129**, 090502 (2022).
 - [4] D. Aharonov, X. Gao, Z. Landau, Y. Liu, and U. Vazirani, A polynomial-time classical algorithm for noisy random circuit sampling, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (ACM Press, New York, NY, 2023), pp. 945–957.
 - [5] C. Oh, M. Liu, Y. Alexeev, B. Fefferman, and L. Jiang, Classical algorithm for simulating experimental Gaussian Boson sampling, *Nat. Phys.* **20**, 1461 (2024).
 - [6] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, *Phys. Rev. Lett.* **119**, 180509 (2017).
 - [7] Y. Li and S. C. Benjamin, Efficient variational quantum simulator incorporating active error minimization, *Phys. Rev. X* **7**, 021050 (2017).
 - [8] A. W. Smith, K. E. Khosla, C. N. Self, and M. Kim, Qubit readout error mitigation with bit-flip averaging, *Sci. Adv.* **7**, eabi8009 (2021).
 - [9] D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, S. Ebadi, M. Cain, M. Kalinowski, D. Hangleiter, *et al.*, Logical quantum processor based on reconfigurable atom arrays, *Nature (London)* **626**, 58 (2024).

- [10] Google Quantum AI and Collaborators, Quantum error correction below the surface code threshold, *Nature (London)* **638**, 920 (2025).
- [11] J. Eisert, D. Hangleiter, N. Walk, I. Roth, D. Markham, R. Parekh, U. Chabaud, and E. Kashefi, Quantum certification and benchmarking, *Nat. Rev. Phys.* **2**, 382 (2020).
- [12] I. L. Chuang and M. A. Nielsen, Prescription for experimental determination of the dynamics of a quantum black box, *J. Mod. Opt.* **44**, 2455 (1997).
- [13] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout, Gate set tomography, *Quantum* **5**, 557 (2021).
- [14] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and robust randomized benchmarking of quantum processes, *Phys. Rev. Lett.* **106**, 180504 (2011).
- [15] R. Harper, S. T. Flammia, and J. J. Wallman, Efficient learning of quantum noise, *Nat. Phys.* **16**, 1184 (2020).
- [16] D. C. McKay, A. W. Cross, C. J. Wood, and J. M. Gambetta, Correlated randomized benchmarking, *arXiv:2003.02354*.
- [17] E. Van Den Berg, Z. K. Mineev, A. Kandala, and K. Temme, Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors, *Nat. Phys.* **19**, 1116 (2023).
- [18] J. Hines, M. Lu, R. K. Naik, A. Hashim, J.-L. Ville, B. Mitchell, J. M. Kriekebaum, D. I. Santiago, S. Seritan, E. Nielsen, *et al.*, Demonstrating scalable randomized benchmarking of universal gate sets, *Phys. Rev. X* **13**, 041030 (2023).
- [19] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Characterizing quantum supremacy in near-term devices, *Nat. Phys.* **14**, 595 (2018).
- [20] D. Hangleiter, M. Kliesch, M. Schwarz, and J. Eisert, Direct certification of a class of quantum simulations, *Quantum Sci. Technol.* **2**, 015004 (2017).
- [21] A. Gheorghiu, T. Kapourniotis, and E. Kashefi, Verification of quantum computation: An overview of existing approaches, *Theory Comput. Syst.* **63**, 715 (2019).
- [22] E. Cruz, F. Baccari, J. Tura, N. Schuch, and J. I. Cirac, Preparation and verification of tensor network states, *Phys. Rev. Res.* **4**, 023161 (2022).
- [23] T. Proctor, S. Seritan, E. Nielsen, K. Rudinger, K. Young, R. Blume-Kohout, and M. Sarovar, Establishing trust in quantum computations, *arXiv:2204.07568*.
- [24] T. Proctor, K. Rudinger, K. Young, E. Nielsen, and R. Blume-Kohout, Measuring the capabilities of quantum computers, *Nat. Phys.* **18**, 75 (2022).
- [25] S. Ferracin, T. Kapourniotis, and A. Datta, Accrediting outputs of noisy intermediate-scale quantum computing devices, *New J. Phys.* **21**, 113038 (2019).
- [26] Z. Wang, Y. Chen, Z. Song, D. Qin, H. Li, Q. Guo, H. Wang, C. Song, and Y. Li, Scalable evaluation of quantum-circuit error loss using Clifford sampling, *Phys. Rev. Lett.* **126**, 080501 (2021).
- [27] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. Van Den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, *et al.*, Evidence for the utility of quantum computing before fault tolerance, *Nature (London)* **618**, 500 (2023).
- [28] D. Gottesman, The Heisenberg representation of quantum computers, *arXiv:quant-ph/9807006*.
- [29] R. Jozsa and A. Miyake, Matchgates and classical simulation of quantum circuits, *Proc. R. Soc. A* **464**, 3089 (2008).
- [30] S. Bravyi and D. Gosset, Improved classical simulation of quantum circuits dominated by Clifford gates, *Phys. Rev. Lett.* **116**, 250501 (2016).
- [31] R. S. Bennink, E. M. Ferragut, T. S. Humble, J. A. Laska, J. J. Nutaro, M. G. Pleszkoch, and R. C. Pooser, Unbiased simulation of near-Clifford quantum circuits, *Phys. Rev. A* **95**, 062337 (2017).
- [32] J. Carrasco, M. Langer, A. Neven, and B. Kraus, Gaining confidence on the correct realization of arbitrary quantum computations, *Phys. Rev. Res.* **6**, L032074 (2024).
- [33] S. Merkel, T. Proctor, S. Ferracin, J. Hines, S. Barron, L. C. G. Govia, and D. McKay, When Clifford benchmarks are sufficient; estimating application performance with scalable proxy circuits, *arXiv:2503.05943*.
- [34] We always assume no single-qubit gate in the layout by simply merging each of them to the nearest two-qubit gate in the circuit.
- [35] L. G. Valiant, Quantum circuits that can be simulated classically in polynomial time, *SIAM J. Comput.* **31**, 1229 (2002).
- [36] B. M. Terhal and D. P. DiVincenzo, Classical simulation of noninteracting-fermion quantum circuits, *Phys. Rev. A* **65**, 032325 (2002).
- [37] B. Bertini, P. Kos, and T. Prosen, Exact correlation functions for dual-unitary lattice models in $1 + 1$ dimensions, *Phys. Rev. Lett.* **123**, 210601 (2019).
- [38] P. Kos and G. Styliaris, Circuits of space and time quantum channels, *Quantum* **7**, 1020 (2023).
- [39] J. I. Cirac, D. Pérez-García, N. Schuch, and F. Verstraete, Matrix product states and projected entangled pair states: Concepts, symmetries, theorems, *Rev. Mod. Phys.* **93**, 045003 (2021).
- [40] B. Kraus and J. I. Cirac, Optimal creation of entanglement using a two-qubit gate, *Phys. Rev. A* **63**, 062309 (2001).
- [41] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, *Phys. Rev. A* **94**, 052325 (2016).
- [42] P. Kos, B. Bertini, and T. Prosen, Correlations in perturbed dual-unitary circuits: Efficient path-integral formula, *Phys. Rev. X* **11**, 011022 (2021).
- [43] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/zn6p-c8w2> for details, which includes Refs. [44,45].
- [44] G. Chiribella, G. M. D'Ariano, and P. Perinotti, Transforming quantum operations: Quantum supermaps, *Europhys. Lett.* **83**, 30004 (2008).
- [45] N. Yoshioka, M. Amico, W. Kirby, P. Jurcevic, A. Dutt, B. Fuller, S. Garion, H. Haas, I. Hamamura, A. Ivrii, R. Majumdar, Z. Mineev, M. Motta, B. Pokharel, P. Rivero, K. Sharma, C. J. Wood, A. Javadi-Abhari, and A. Mezzacapo, Krylov diagonalization of large many-body Hamiltonians on a quantum processor, *Nat. Commun.* **16**, 5014 (2025).
- [46] S. P. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).
- [47] S. T. Flammia and R. O'Donnell, Pauli error estimation via population recovery, *Quantum* **5**, 549 (2021).
- [48] C. Rouzé and D. S. Franca, Efficient learning of the structure and parameters of local Pauli noise channels, *arXiv:2307.02959*.

- [49] F. Baccari, P. Kos, and G. Styliaris, Average-computation benchmarking for local expectation values in digital quantum devices, Zenodo (2025), <https://doi.org/10.5281/zenodo.16731873>.
- [50] L. E. Fischer, M. Leahy, A. Eddins, N. Keenan, D. Ferracin, M. A. C. Rossi, Y. Kim, A. He, F. Pietracaprina, B. Sokolov, S. Dooley, Z. Zimborás, F. Tacchino, S. Maniscalco, J. Goold, G. García-Pérez, I. Tavernelli, A. Kandala, and S. N. Filippov, Dynamical simulations of many-body quantum chaos on a quantum computer, *Nat. Phys.* **22**, 302 (2026).
- [51] For other positions, the expectation values vanish.
- [52] R. M. Milbradt, L. Scheller, C. Aßmus, and C. B. Mendl, Ternary unitary quantum lattice models and circuits in $2 + 1$ dimensions, *Phys. Rev. Lett.* **130**, 090601 (2023).
- [53] R. Suzuki, K. Mitarai, and K. Fujii, Computational power of one- and two-dimensional dual-unitary quantum circuits, *Quantum* **6**, 631 (2022).
- [54] C. Jonay, V. Khemani, and M. Ippoliti, Triunitary quantum circuits, *Phys. Rev. Res.* **3**, 043046 (2021).