Circuit connectivity boosts by quantum-classical-quantum interfaces

Roeland Wiersema,[1,2,*] Leonardo Guerini,[3,4] Juan Felipe Carrasquilla,[1,2,5] and Leandro Aolita[4,6]
[1]*Vector Institute, MaRS Centre, Toronto, Ontario, Canada M5G 1M1*
[2]*Department of Physics and Astronomy, University of Waterloo, Ontario, Canada N2L 3G1*
[3]*Department of Mathematics, Federal University of Santa Maria, Santa Maria, Rio Grande do Sul 97105-900, Brazil*
[4]*Instituto de Física, Federal University of Rio de Janeiro, P.O. Box 68528, Rio de Janeiro 21941-972, Brazil*
[5]*Department of Physics, University of Toronto, Ontario, Canada M5S 1A7*
[6]*Quantum Research Centre, Technology Innovation Institute, Abu Dhabi, United Arab Emirates*

(Received 26 April 2022; accepted 5 December 2022; published 29 December 2022)

High-connectivity circuits are a major roadblock for current quantum hardware. We propose a hybrid classical-quantum algorithm to simulate such circuits without SWAP-gate ladders. As the main technical tool, we introduce quantum-classical-quantum interfaces. These replace an experimentally problematic gate (e.g., a long-range one) with single-qubit random measurements followed by state preparations sampled according to a classical quasiprobability simulation of the noiseless gate. Each interface introduces a multiplicative statistical overhead which, remarkably, is independent of the on-chip qubit distance. Hence, by applying interfaces to the longest-range gates in a target circuit, significant reductions in circuit depth and gate infidelity can be attained. We numerically show the efficacy of our method for a Bell-state circuit for two increasingly distant qubits and a variational ground-state solver for the transverse-field Ising model on a ring. Our findings provide a versatile toolbox for error-mitigation and circuit boosts tailored for noisy, intermediate-scale quantum computation.

DOI: 10.1103/PhysRevResearch.4.043221

I. INTRODUCTION

Quantum computation promises a major disruption in high-performance computing, with applications in diverse fields ranging from many-body physics and chemistry to machine learning, finance, automation, or logistics, to name a few [1–3]. However, the current paradigm of noisy, intermediate-scale quantum (NISQ) devices limits quantum algorithms to circuits with low qubit numbers, low depth, and low connectivity [4]. This poses serious concerns regarding the actual usefulness of quantum computers in the near term and has thus ignited a both experimental and theoretical quest for ways to unleash the potential of quantum algorithms with NISQ hardware [5–7].

A large class of NISQ algorithms are based on hybrid quantum-classical approaches. One of the most successful of these consists of parametrized quantum circuits variationally optimized through a classical optimizer aimed at approximating a target ground state [8,9]. To combat the noise in these systems, subsequent variants incorporated the idea of quantum error mitigation [10–13]. This refers to schemes whereby noisy experimental implementations (e.g., in different noise regimes or with different gate choices), together with suitable classical postprocessing, are used to simulate a target, noiseless quantum circuit of limited size. This offers a NISQ alternative to quantum error correction (which requires large-scale quantum circuits), where full fault tolerance is achieved by actively correcting errors on the quantum hardware during the execution of the computation.

More recently, a different type of hybrid method has been put forward [14–19]. There, a classical algorithm calls a quantum computer as a subroutine to simulate a larger quantum circuit. However, the cost of this is that both the number of queries to the quantum subroutine and the classical post-processing runtime unavoidably grow exponentially with the size of the target circuit. Moreover, a particularly challenging aspect of NISQ devices is their inability to run algorithms that require high, long-range connectivity among the constituent qubits. In most NISQ hardware, long-range gates are synthesized by a long sequence of nearest-neighbor gates. This drastically inflates the circuit depth and causes large infidelity due to noise accumulation incurred during the syntheses. This is a crucial limitation in the NISQ era.

Here, we take a conceptually different direction from previous hybrid schemes: Instead of assembling a large quantum circuit from small pieces, we simulate a high-connectivity circuit from circuits with low connectivity and depth. To that end, we introduce the notion of *quantum-classical-quantum* (QCQ) interfaces. A QCQ interface for a gate U corresponds to a local measurement on the qubits on which U acts followed by a repreparation of those same qubits in a random product state that depends on U. In other words, the interface performs a hybrid quantum-classical simulation of U. Each interface introduces a multiplicative statistical overhead that,

*rwiersema@uwaterloo.ca

as we prove below, is independent of the on-chip distance between the qubits. Hence, for a fixed number of interfaces, for example, the longer the range of the target gates is, the more drastic the reduction in depth attained is at the expense of a constant overall statistical overhead.

More technically, our interfaces combine state-of-the-art state estimation based on single-qubit random measurements [20,21] with quasiprobability representations based on frames [22,23]. Such representations have been used for classically simulating a quantum circuit with Monte Carlo sampling techniques [24–26]. In particular, our algorithm can be seen as a hybrid version of the scheme of Ref. [26], where everything is quantum except for a subset of gates that one wishes to "cut out" of the experimental circuit. Here, we choose such a subset in terms of the on-chip qubit distance. However, other relevant choices may be due simply to error mitigation or hardware-specific limitations. Like most quasiprobability schemes, our method suffers from the infamous sign problem [27–29]. Remarkably, the severity of the problem depends only on the number of interfaces and not on the on-chip distance between the qubits. Moreover, as a by-product contribution, in order to minimize the average sign of our quasiprobability representation, we develop a Metropolis-Hastings simulated-annealing algorithm based on random walks in the space of dual positive operator-valued measurements (POVMs). We implement such walks through a convenient, long-known parametrization of generalized inverse matrices [30]. This allows us to decrease the sample-complexity overhead per interface by almost a factor of 4 relative to the canonical POVM choice, constituting a practical tool of general relevance for sign-problem mitigation [31,32].

The paper is organized as follows. In Sec. II we introduce our notation and the necessary mathematical background to understand QCQ interfaces. We then present our algorithm in detail in Sec. III. In Sec. IV we perform numerical experiments to show the efficacy of our method on two illustrative circuits, namely, the preparation of a Bell state between two increasingly distant qubits and a variational ground-state solver for the one-dimensional (1D) transverse-field Ising model with periodic boundary conditions. We end with a discussion of our results in Sec. V and provide a perspective on other potential applications of our method.

II. PRELIMINARIES

Here, we give a high-level description of our method and leave the formal treatment in terms of frame theory [22,23] for Appendix A.

We consider an N-qubit system $\mathfrak{f}\mathcal{S}$ described by a density matrix ϱ. This density matrix can be fully described via the measurement statistics of an informationally complete positive operator-valued measure (IC-POVM) $\boldsymbol{M} = \{M_{\boldsymbol{a}}\}_{\boldsymbol{a}\in\{1,\dots,m\}^N}$, which can be constructed by taking the tensor product of single-qubit IC-POVMs, $M_{\boldsymbol{a}} = M_{a_1}\otimes\cdots\otimes M_{a_N}$, where $M_{a_i} \geqslant 0$ and $\sum_{a_i=1}^{m} M_{a_i} = \mathbb{1}$ [33,34]. For each operator $M_{\boldsymbol{a}}$ we can define a dual IC-POVM element $\tilde{M}_{\boldsymbol{a}}$ such that the following equality holds:

$$\varrho = \sum_{\boldsymbol{a}} P_{\varrho}(\boldsymbol{a})\tilde{M}_{\boldsymbol{a}}, \tag{1}$$

where $P_{\varrho}(\boldsymbol{a}) := \mathrm{Tr}[M_{\boldsymbol{a}}\varrho]$ is the probability of measurement outcome $\boldsymbol{a}$ on ϱ. Equation (1) is the basis of classical-shadow tomography, a powerful technique to get compact classical representations of states from measurements [21,35]. Note that Eq. (1) also works if $M_{\boldsymbol{a}}$ acts on a subset of all N qubits, e.g., $\varrho = \sum_{\boldsymbol{a}} P_{\varrho}(\boldsymbol{a})(\tilde{M}_{\boldsymbol{a}}\otimes\varrho_{\mathrm{red}}(\boldsymbol{a}))$, where $\varrho_{\mathrm{red}}(\boldsymbol{a})$ is the normalized state on the rest of the system after applying $M_{\boldsymbol{a}}$.

The dual POVM elements $\tilde{M}_{\boldsymbol{a}}$ can be expressed in terms of the $M_{\boldsymbol{b}}$'s as

$$\tilde{M}_{\boldsymbol{a}} = \sum_{\boldsymbol{b}} \tilde{T}_{\boldsymbol{a},\boldsymbol{b}} M_{\boldsymbol{b}}, \tag{2}$$

where $\tilde{T} = \mathfrak{T}\,T\,\mathfrak{T}$ and $T_{\boldsymbol{a},\boldsymbol{b}} := \mathrm{Tr}[M_{\boldsymbol{a}}M_{\boldsymbol{b}}]$; hence T is an $m^N \times m^N$ matrix. The matrix $\mathfrak{T}$ has to satisfy the equation $T = T\mathfrak{T}T$, i.e., $\mathfrak{T}$ is a generalized inverse of T [36].

By virtue of Eqs. (1) and (2), we can then express any ϱ as an affine combination of product states by normalizing the POVM elements.

This fact has been used to reconstruct quantum states [20], processes [37], and overlaps [38] from single-qubit measurements. Additionally, this has been used to simulate quantum circuits [39] with generative machine learning models, where $\mathfrak{T}$ was taken as the canonical pseudoinverse of T. However, other choices of $\mathfrak{T}$ are possible. The columns of $\mathfrak{T}$ are normalized, but in general, its elements can be positive or negative; hence we can understand it as a quasiprobability distribution [26]. The negativity of $\mathfrak{T}$ has important consequences for the sample complexity of our algorithm.

III. INTERFACES FOR HYBRID CLASSICAL-QUANTUM CIRCUITS

Our goal is to simulate observable measurements on quantum circuits using hybrid classical-quantum ones. More precisely, we are given an observable O, an N-qubit input state $\varrho_0 := |0\rangle\langle 0|$, and a target circuit $C := \{U_k\}_{k=1,\dots,K}$, consisting of single- or two-qubit unitary gates U_k. We denote by $s_k \subset \mathcal{S}$ the subset of qubits on which U_k acts, and by $\boldsymbol{a}_{s_k}$ a corresponding substring of POVM measurement outcomes on s_k. Some of the gates $\{U_l\}_{l=l_1,\dots,l_L} \subset C$, where $l_i \leqslant K$ are gates in the circuit that we intend to replace, because they are, for instance, particularly experimentally demanding for NISQ implementations or do not match the native hardware connectivity of the device.

The case we explicitly study below is that of two-qubit gates on qubits far apart in the connectivity graph in question. We want to estimate the expectation value $\mathrm{Tr}[\varrho_K O]$ of O on the output state $\varrho_K := U_K\cdots U_1\,\varrho_0\,U_1^{\dagger}\cdots U_K^{\dagger}$ by substituting every U_k with $k \in L$ by a classical simulation of it. Our main tool to achieve this consists of interfaces between quantum objects and their (classical) representations. Note that we can also use a partial state reconstruction. The first type of interface is based on Eq. (1).

Definition 1: Quantum-classical interfaces. A quantum-classical (QC) interface on s_k refers to the assignment of a classical snapshot $\tilde{M}_{\boldsymbol{a}_{s_k}}$ according to the measurement outcome $\boldsymbol{a}_{s_k}$ of a factorable POVM $\boldsymbol{M}$ on a state ϱ, occurring with probability $P_{\varrho}(\boldsymbol{a}_{s_k}) := \mathrm{Tr}[M_{\boldsymbol{a}_{s_k}}\varrho]$.

Hence the QC interface is a list of bit strings corresponding to measurement outcomes on the subsystem s_k.

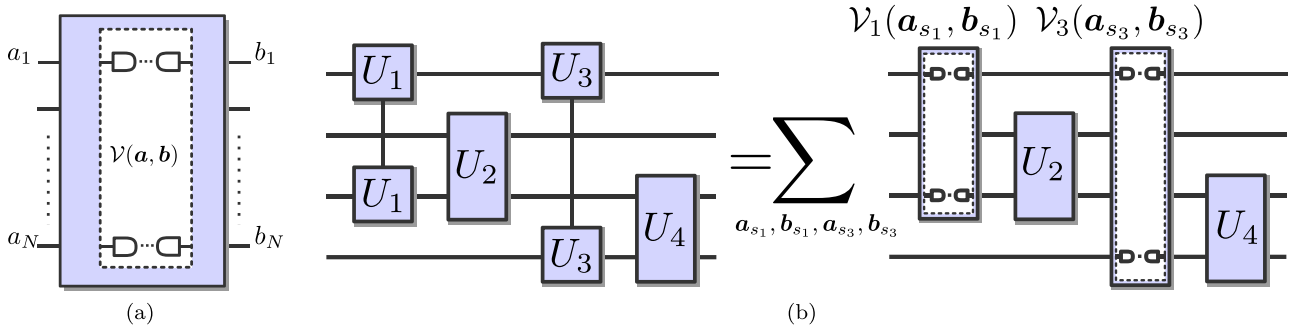

FIG. 1. Schematics of our method. (a) A QCQ interface $\mathcal{V}(\boldsymbol{a},\boldsymbol{b})$ applying the identity operator $U_k = \mathbb{1}$ between qubits 1 and N. We measure the POVM $\boldsymbol{M}$ on both qubits reprepare them in a product state that depends on the simulated gate and the outcome $\boldsymbol{a}$. The other $N-2$ qubits are left untouched. (b) An exemplary four-qubit circuit (left) is simulated by a hybrid quantum-classical circuit (right), where the non-nearest-neighbor gates $\mathcal{U}_1$ and $\mathcal{U}_3$ are substituted by QCQ interfaces $[\mathcal{V}_1(\boldsymbol{a}_{s_1},\boldsymbol{b}_{s_1})$ and $\mathcal{V}_3(\boldsymbol{a}_{s_3},\boldsymbol{b}_{s_3})$, respectively]. The summation over $(\boldsymbol{a}_{s_1},\boldsymbol{b}_{s_1},\boldsymbol{a}_{s_3},\boldsymbol{b}_{s_3})$ represents the average over all interface outcomes sampled (see text).

From the outcomes $\boldsymbol{a}_{s_k}$ of the QC interface, we can use Eq. (2) to reconstruct the state by importance-sampling $\boldsymbol{b}_{s_k}$. To achieve this, we first define the normalized states $\sigma_{\boldsymbol{b}} := M_{\boldsymbol{b}}/\mathrm{Tr}[M_{\boldsymbol{b}}]$. Next, we rewrite $\tilde{T}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$ as

$$\tilde{T}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} =: \left\|\tilde{T}_{\boldsymbol{a}_{s_k}}\right\|_1 \mathrm{sgn}\big(\tilde{T}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}\big)\, P\big(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k}\big), \tag{3}$$

where $\|\tilde{T}_{\boldsymbol{a}_{s_k}}\|_1 := \sum_{\boldsymbol{b}_{s_k}} |\tilde{T}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}|$ is the l_1 norm of the rows $\tilde{T}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$ and $P(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k}) := |\tilde{T}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}|/\|\tilde{T}_{\boldsymbol{a}_{s_k}}\|_1$ is the conditional probability distribution obtained by taking the absolute value of the rows and normalizing appropriately. By construction, $P(\circ|\boldsymbol{a}_{s_k})$ is a valid probability distribution, which allows us to quantum-Monte-Carlo-simulate $\tilde{M}_{\boldsymbol{a}_{s_k}}$ by sampling $\boldsymbol{b}_{s_k}$ [26]. This leads us to the definition of our second type of interface.

Definition 2: Classical-quantum interface. A classical-quantum (CQ) interface on s_k refers to the repreparation of the state $\sigma_{\boldsymbol{b}_{s_k}}$, with probability $P(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k})$, given a classical snapshot $\tilde{M}_{\boldsymbol{a}_{s_k}}$. Each sampled pair $(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k})$ is assigned the value $\|\tilde{T}_{\boldsymbol{a}_{s_k}}\|_1 \,\mathrm{sgn}(\tilde{T}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}})\mathrm{Tr}[M_{\boldsymbol{b}}]$.

The CQ interface is thus a collection of bit strings indicating which state to reprepare on s_k, while we keep track of the signs and norms of $\tilde{T}$. One can combine the QC interface with the CQ interface to represent ϱ by measuring and repreparing states on s_k. The main contribution of our work is going beyond this identity. To do this, we absorb the action of a gate U_k acting on s_k into the measurement and repreparation of ϱ by defining $\tilde{T}^{U_k} := \mathfrak{T}\, T^{U_k}\, \mathfrak{T}$, where

$$T^{U_k} := \mathrm{Tr}\big[U_k M_{\boldsymbol{a}_{s_k}} U_k^\dagger M_{\boldsymbol{b}_{s_k}}\big]. \tag{4}$$

We provide a derivation of this quantity in Appendix A. This leads us to our final definition.

Definition 3: Quantum-classical-quantum interface. A quantum-classical-quantum (QCQ) interface on s_k given a gate U_k refers to the measurement of $\boldsymbol{M}$ with outcome $\boldsymbol{a}_{s_k}$, followed by the repreparation of $\sigma_{\boldsymbol{b}_{s_k}}$ with probability $P_{U_k}(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k})$. Each sampled pair $(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k})$ is assigned the value $v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} := \|\tilde{T}^{U_k}_{\boldsymbol{a}_{s_k}}\|_1 \,\mathrm{sgn}(\tilde{T}^{U_k}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}})\mathrm{Tr}[M_{\boldsymbol{b}}]$. We represent this interface by $\mathcal{V}_k(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k})$.

Note that we can place an interface at any point in the circuit to replace a gate. For example, we can perform the gates $\{U_1\cdots U_{l_1-1}\}$ to our initial state ϱ_0, create an interface $\mathcal{V}_{l_1}(\boldsymbol{a}_{s_{l_1}},\boldsymbol{b}_{s_{l_1}})$, and then apply the rest of the circuit $\{U_{l_1+1}\cdots U_K\}$ to the reprepared state $\sigma_{s_{l_1}}$ (Fig. 1). By combining Eqs. (1)–(3) we can obtain the following equation for the expectation value of an observable O via a QCQ interface:

$$\mathrm{Tr}[\varrho O] = \sum_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} P_\varrho\big(\boldsymbol{a}_{s_{l_1}}\big)\, v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} \mathrm{Tr}\big[M_{\boldsymbol{b}_{s_k}}\big] \times \mathrm{Tr}\big[U_K\cdots U_{l_1+1}\sigma_{\boldsymbol{b}_{s_k}} U_{l_1+1}\cdots U_K O\big]. \tag{5}$$

We can extend the single-QCQ-interface example above to multiple interfaces by applying subsequent measurement-and-repreparation steps and multiplying the norms and signs $v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$ of each interface accordingly.

Equation (5) and its generalization to multiple interfaces can be experimentally calculated with a finite-statistics estimator $O^*_{M_s}$ over M_s runs (see Appendix D). We refer to M_s as the *sample complexity* of our protocol. Clearly, the estimation of observables via QCQ interfaces comes at a cost. In particular, the multiplicative factors of $v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$ increase the variance of the observable estimator $O^*_{M_s}$; hence we need more runs M_s to get an accurate estimate of $\mathrm{Tr}[\varrho_K O]$. In practice, M_s needs to be chosen to guarantee that the statistical error and significance level (failure probability) of the estimation are given by target values ε and δ, respectively. The entire procedure is sketched by the pseudocode in Algorithm 1.

To quantify the runtime of the algorithm given ε and δ, we define the *interface negativity* of the gate U_k and the *total forward interface negativity* of the entire circuit C as

$$n_{U_k} := \max_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} \left\|\tilde{T}^{U_{s_k}}_{\boldsymbol{a}_{s_k}}\right\|_1 \mathrm{Tr}\big[M_{\boldsymbol{b}_{s_k}}\big] \quad \text{and} \quad n_\rightarrow := \prod_{k\in L} n_{U_k}, \tag{6}$$

respectively. This allows us to state the following theorem.

Theorem 1: Correctness and sample complexity. The finite-statistics average O^*_M of Algorithm 1 is an unbiased estimator of $\mathrm{Tr}[\varrho_K O]$ (see Appendix D). Moreover, if

$$M_s \geqslant n_\rightarrow^2 \times \frac{2\,\|O\|^2\,\ln(2/\delta)}{\varepsilon^2}, \tag{7}$$

with $\|O\|$ being the operator norm of O, then, with probability at least $1-\delta$, the statistical error of O^*_M is at most ε.

The proof follows straightforwardly from the Hoeffding bound. We note that the factor $\frac{2\,\|O\|^2\,\ln(2/\delta)}{\varepsilon^2}$ in Eq. (7) is

Algorithm 1. Hybrid classical-quantum simulation with QCQ interfaces.

Input: $\varrho_0, C, O, \varepsilon, \delta$
Output: $O^*_{M_s}$ s.t. $|O^*_{M_s} - \mathrm{Tr}[O\,\varrho_K]| \leqslant \varepsilon$ with probability at least $1-\delta$.
Initialize $O^*_{M_s} = 0$, $v = 1$, and M_s as in Eq. (7).
for $m \in (1, \ldots, M_s)$ **do**
  for $k \in (1, \ldots, K)$ **do**
    if $k \in \{l_1, \ldots, l_L\}$ **them**
      Apply a QCQ interface for U_k on qubits s_k, obtaining the pair $(\boldsymbol{a}_{s_k}, \boldsymbol{b}_{s_k})$;
      $v \leftarrow v \times v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$, with $v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$ as in Definition 3.
    else
      Apply the gate U_k on qubits s_k.
    end
  end
  Measure O, obtaining the measurement outcome (eigenvalue of O) o;
  $O^*_{M_s} \leftarrow O^*_{M_s} + o \times v$.
end
$O^*_{M_s} \leftarrow \frac{O^*_{M_s}}{M_s}$.

the equivalent sample-complexity bound one would obtain if $\mathrm{Tr}[\varrho_K\, O]$ was estimated from measurements on the actual state ϱ_K. Hence $n^2_{\rightarrow}$ quantifies the runtime overhead introduced by the interfaces. In that regard, the interface negativities play the same role in our hybrid classical-quantum simulation as the negativities of Ref. [26] in fully classical simulations with quasiprobability representations. An innovative and advantageous feature of Eq. (6) is the presence of the POVM-element trace $\mathrm{Tr}[M_{\boldsymbol{b}_{s_k}}]$ in n_{U_k}, which comes from the state repreparation. Indeed, since $\mathrm{Tr}[M_{\boldsymbol{b}_{s_k}}] < 1$, the n_{U_k}'s (and therefore also $n_{\rightarrow}$) are significantly smaller than their counterparts for fully classical simulations [26]. This is consistent with the intuition that hybrid classical-quantum Monte Carlo simulations should cause lower sample-complexity increases than fully classical ones. Our bound is similar to the sample complexity of the spacelike circuit cuts in Refs. [19,40] but is not restricted specific gates.

Either way, the most relevant property for our purposes is that $n^2_{\rightarrow}$ (and therefore also M_s) is independent of not only the number of gates K or qubits N but also, most importantly, the connectivity-graph distance between the qubits on which the interfaces act. In other words, for a fixed budget of measurement runs, simulating a gate U_k with a QCQ interface increases the statistical error at most by a constant factor n_{U_k}, regardless of how far apart in the circuit the qubits s_k are. In contrast, experimentally synthesizing U_k with noisy nearest-neighbor gates would give a systematic error due to infidelity accumulation that grows with the distance between those qubits.

With regard to the limitations of our method, we note that $n^2_{\rightarrow}$ grows exponentially with the number L of interfaces used. We can therefore only simulate a limited a number of gates before the number of measurement-and-repreparation steps becomes too large to perform in practice. Additionally, the forward negativity depends on $\|\tilde{T}^{U_{s_k}}_{\boldsymbol{a}_{s_k}}\|_1$, which increases with the number of qubits onto which the simulated gate acts. However, we are usually only interested in simulating two-qubit gates, where this effect is small. Even with these drawbacks, Algorithm 1 constitutes a better alternative for many circuits than the bare NISQ implementation. Also, Theorem 1 provides a direct way to get a sense of whether implementing a QCQ interface will be too difficult to perform in practice, since we can obtain an upper bound on the number of shots required to perform an accurate simulation of a certain gate. We study relevant exemplary circuits with such trade-offs in the next sections.

Finally, note that $n^2_{\rightarrow}$ is POVM dependent. This is crucial to the efficiency of classical simulations [27–29]. For instance, in the quantum Monte Carlo method, it is known that the statistical overhead due to negative (quasi)probabilities can be ameliorated [32] or even removed [31] by local base changes. Something similar applies here: The interface negativities depend not only on the choice of POVM, but also on how we construct the dual POVM elements.

IV. NUMERICAL EXPERIMENTS

Here, we provide numerical experiments to validate the procedure outlined in Algorithm 1. Throughout the rest of this paper, we take $\{M_{\boldsymbol{a}}\}_{\boldsymbol{a}}$ to be the Pauli-6 IC-POVM,

$$\{M_{\boldsymbol{a}}\}^{\text{Pauli-6}}_{\boldsymbol{a}} := \bigcup_{i=x,y,z} \left\{ \frac{1}{3}|\uparrow_i\rangle\langle\uparrow_i|, \frac{1}{3}|\downarrow_i\rangle\langle\downarrow_i| \right\}, \tag{8}$$

where the vectors $|\uparrow_i\rangle$ and $|\downarrow_i\rangle$ correspond to the eigenvectors of the Pauli operators with eigenvalues $+1$ and -1, respectively. Note that this POVM can be implemented in an experimental setting without the usage of ancilla qubits (see Appendix B).

For our simulations, we make use of full density matrix simulations and locally purified density operator (LPDO) tensor networks [41] (see Appendix F). For the latter, we choose the bond and Kraus dimensions D and κ, respectively, such that the simulation errors are under control and we end up with a high-fidelity (>99.9%) state approximation. To simulate realistic experimental settings, we apply noise to the two-qubit gates in our circuit. In particular, we implement noisy CNOT gates throughout our circuits by applying single-qubit depolarizing channels $\mathcal{E} : \varrho \mapsto \mathcal{E}(\varrho)$ to both the control and target qubit of the CNOT gate. We apply depolarizing noise in the CNOT gates with $\lambda_{\text{unit}} = 0.005$. These values correspond to experimentally realistic values [42]. At the end of the circuit we estimate observables $\mathrm{Tr}[\varrho O]$ exactly, i.e., without further sampling bit strings but relying on the full state representation.

Since we are considering two-qubit gates for our numerical experiments, our interfaces only act on two-qubit systems. Hence, for our measurement-and-repreparation step, we only need to store bit strings $\boldsymbol{a}_{s_l}$ of length 2, as well as the $6^2 \times 6^2$ overlap matrix $\tilde{T}^{U_k}$.

To improve the sample complexity of our algorithm, we use a Monte Carlo algorithm to minimize the interface negativities. We first note that the matrix $\tilde{T}^{U_k}$ defined under Eq. (2) defines a domain over which to optimize such negativity. Similar optimizations have been used for alleviating the sign problem in partition-function estimations [31,32]. In our setting, we use a convenient parametrization of generalized

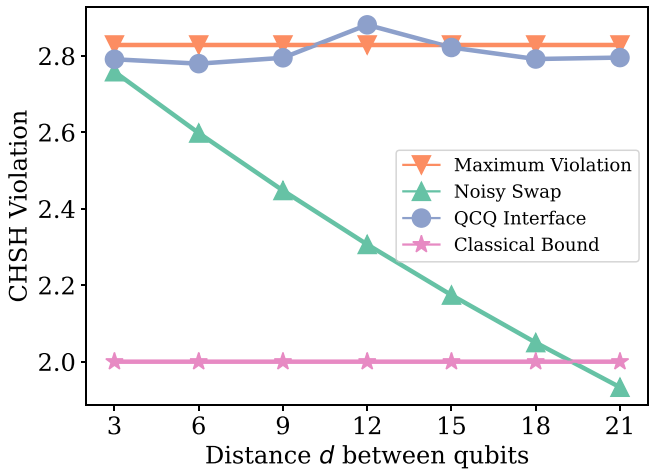

FIG. 2. CHSH violation as a function of the number of qubits. These results were obtained with an LPDO simulation where $D = 12$ and $\kappa = 24$. In addition to the gate noise, we apply a depolarizing channel to simulate measurement noise with $\lambda_{\text{meas}} = 0.01$ and repreparation noise with $\lambda_{\text{reprep}} = 0.005$. The classical bound (pink) and maximal violation orange) are 2 and $2\sqrt{2}$, respectively, for all d. We see that the violation in the noisy circuit (green) decreases linearly with the number of qubits as a result of the $4(d-2)+1$ noisy SWAP gates required to prepare the state. Our algorithm provides the maximum CHSH violation up to statistical fluctuations independent of the distance between the qubits. This comes at a cost of sampling $M = 60\,000$ measurement-and-repreparation steps to estimate the violation.

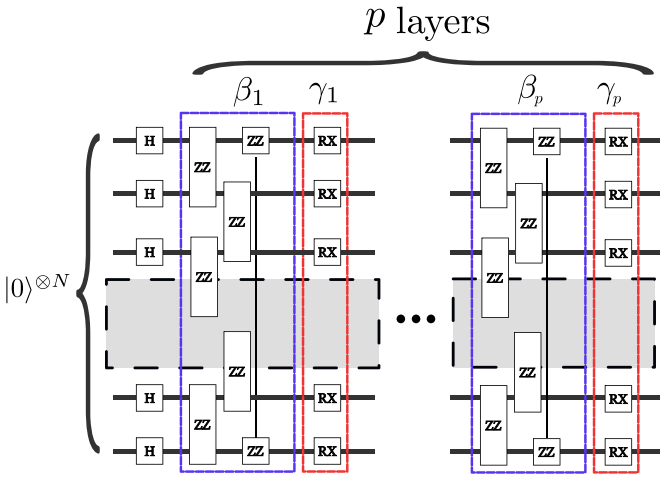

FIG. 3. The Hamiltonian variational ansatz circuit for the ground state of the TFIM. The parameters $\{\beta_i, \gamma_i\}$ for $i = 1, \ldots, p$ can be found with a variational quantum eigensolver optimization [9]. Each layer in the circuit contains a long-range two-qubit ZZ rotation. We assume that the distance between the first and last qubit is $N-2$. Implementing the nearest-neighbor ZZ gates comes at a cost of $2(N-1)$ CNOT gates. The long-range ZZ rotations require $4(N-2)+1$ CNOT gates since we must use a SWAP chain to bring the first and last qubit together. The total number of CNOT gates per layer is therefore dominated by the implementation of long-range ZZ rotations.

inverse matrices by Rao [30] to propose dual POVM elements for an adaptive random walk Metropolis-Hastings algorithm. This allows us to decrease the multiplicative sample-complexity overhead per interface by almost a factor of 4 relative to the canonical dual POVM (corresponding to $\tilde{T} = T^{-1}$, with T^{-1} being the pseudoinverse of T), which reduces the number of samples required by a factor of 4 (see Appendix G).

A. Simulation of long-range maximal Bell violations

As a proof-of-principle experiment, we show that a maximally entangled state simulated with our method attains the maximal violation of the Clauser-Horne-Shimony-Holt (CHSH) inequalities (see Appendix E) as expected. Specifically, we create the Bell state $|\Phi^+\rangle = \frac{1}{2}(|00\rangle + |11\rangle)$, which has the maximum CHSH violation $S(A, B) = 2\sqrt{2}$. We consider the case where the state is prepared on two qubits separated by a distance d. Applying the CNOT between these distant qubits requires implementing a SWAP chain to bring the two states close together. In Fig. 2 we compare the CHSH violation of the Bell state simulated with our algorithm and one prepared with a circuit containing a noisy SWAP chain. We see that the CHSH violation is only affected by the statistical fluctuations of our method and therefore approximates the maximum value independent of the distance between the qubits.

B. The transverse-field Ising-model circuit

As a practical example of implementing our method in an experimentally realistic setting, we investigate the ground state of a prototypical model for quantum magnetism: the transverse-field Ising model (TFIM) on a one-dimensional ring. The Hamiltonian of the TFIM for the 1D chain is given by

$$H_{\text{TFIM}} = -\sum_{i=1}^{N} [Z_i Z_{i+1} + g X_i], \tag{9}$$

where we assume periodic boundary conditions and set $g = 1$. The ground state of H can be approximated reliably with a depth $p = N/2$ circuit ansatz called the Hamiltonian variational ansatz [43–45].

This circuit for the ground state is given in Fig. 3. To evaluate the accuracy of the state reconstruction, we compare the finite-statistics estimator of the energy $\langle \hat{H}_M \rangle$ from our algorithm with the ground-state energy $E_{\text{gs}} = \langle \psi_{\text{gs}} | H | \psi_{\text{gs}} \rangle$ from exact diagonalization.

We consider three setups: First, we consider the $N = 4$ and $N = 8$ qubit TFIM chains where the last long-range ZZ gate (in the second and fourth layers, respectively) is classically simulated with our algorithm (see Fig. 4). Next, we apply our method twice for the same circuits, with simulation of both the last and first-to-last long-range ZZ gate (see Fig. 5). Finally, we consider the ground state of an $N = 20$ TFIM chain, where we only apply the first two layers of the circuit and simulate the second long-range ZZ gate (see Fig. 6). For all experiments, we confirm that we can greatly improve the final energy estimates by making use of QCQ interfaces at the cost of M_s measurement-and-repreparation steps.

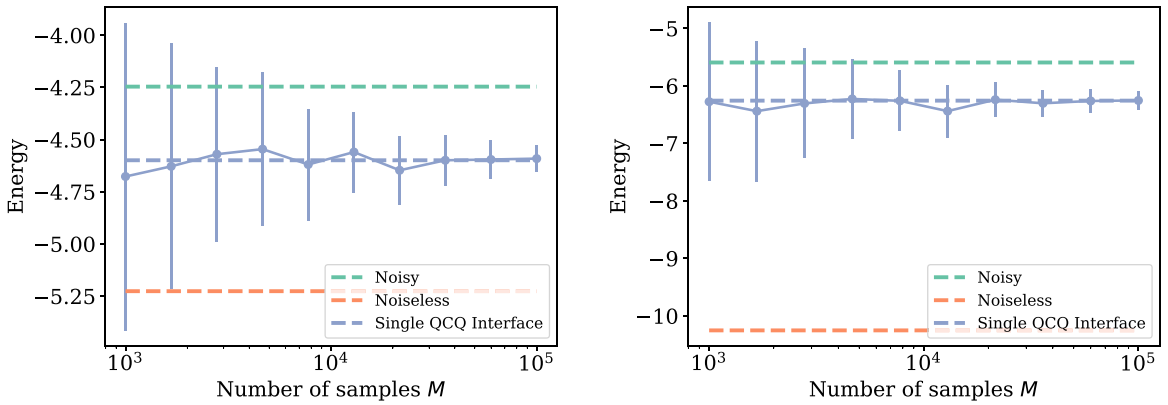

FIG. 4. Comparison of QCQ interface simulation with both noisy and noiseless TFIM circuits for (a) $N = 4$ and (b) $N = 8$ qubits obtained with a full density matrix simulation. Each dot represents the average energy $\mathbb{E}[\langle \hat{H}_M \rangle]$ estimated over 50 separate instances. The error bars indicate the standard deviation. As the number of samples M_s increases, the statistical fluctuations of our method become small in accordance with the central limit theorem. We can determine the scaling of the size of the error bars by fitting $\sigma = \bar{\sigma}/\sqrt{N_{\text{samples}}}$. While for four qubits $\bar{\sigma} \approx 27.8$, for eight qubits we have $\bar{\sigma} \approx 76.5$. This scaling only depends on the mean negativity, which differs between the two circuits because we apply a different ZZ rotation on each circuit. The energy of the noiseless circuit (orange dashed line) corresponds to the ground-state energy E_{gs}. The noisy circuit (green dashed line) shows the energy obtained when we apply depolarizing channels with $\lambda_{\text{unit}} = 0.005$ to the CNOT gates in the circuit. We see that for both four and eight qubits, our algorithm provides a significant improvement on the final estimated energy of the circuit for a reasonable number of measure-and-reprepare steps. In (b) we observe that the large number of number of noisy CNOT gates dominates the simulation; hence the improvement is not as significant as for four qubits.

V. FINAL DISCUSSION

We have introduced a rigorous framework of hybrid quantum-classical interfaces for quantum-circuit simulations. We applied a specific variant of these gadgets—which we dub quantum-classical-quantum (QCQ) interfaces—to simulate long-range gates in low-connectivity devices without using SWAP-gate ladders. QCQ interfaces replace an experimentally problematic gate (e.g., a very long range one) by single-qubit random measurements and state preparations sampled according to a classical quasiprobability simulation of the ideal target gate. This procedure eliminates long SWAP-gate ladders which would otherwise be required to physically synthesize the target gate. This results in a drastic increase in gate fidelity. The final output of the scheme is an estimate of the expectation value of a given observable on the output of the target high-connectivity circuit.

The quasiprobability distribution used is given by a POVM representation of the gate simulated at each interface. As with any sampling scheme based on nonpositive quasiprobabilities, our method suffers from the sign problem. Because of this, the overall sample complexity grows exponentially with the number of interfaces applied. However, the statistical

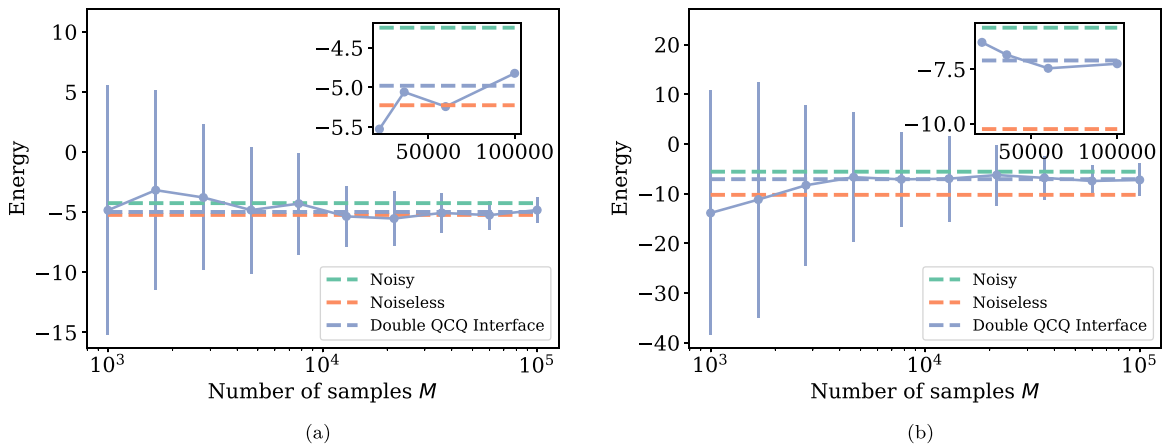

FIG. 5. Comparison of double-QCQ-interface simulation with both noisy and noiseless TFIM circuits for (a) $N = 4$ and (b) $N = 8$ qubits. These results were obtained with a full density matrix simulation. In (a), we see that we can almost approximate the true ground-state energy of the four-qubit state, because the only noisy operations are the 12 CNOT gates required for implementing the six nearest-neighbor ZZ gates in layers 1 and 2. In (b) we see a more significant improvement over the energies from Fig. 4(b), but still the noise dominates. Since we apply the QCQ method twice, the standard deviation $\sigma = \bar{\sigma}/\sqrt{N_{\text{samples}}}$ of the error bars increases quadratically, as per Eq. (6). We find $\bar{\sigma} \approx 333.1$ and $\bar{\sigma} \approx 856.8$ for four and eight qubits, respectively.

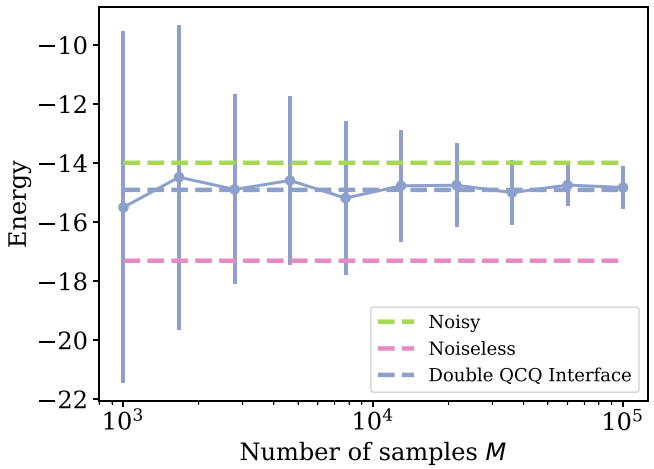

FIG. 6. Comparison of a QCQ interface simulation with both noisy and noiseless circuits for a 20-qubit TFIM circuit. These results were obtained with an LPDO simulation where $D = 50$ and $\kappa = 50$. Only two of the eight layers of the circuit are simulated here, to keep simulation errors under control. The sample variance $\hat{\sigma} \approx 195.0$.

overhead per interface is independent of the on-chip distance between the qubits on which the interface acts. To ameliorate the sign problem, we developed a Metropolis-Hastings simulated-annealing algorithm based on random walks in the space of dual POVMs. This allowed us to decrease the statistical overhead per interface by almost a factor of 2 over that of the canonical dual POVM. This is potentially interesting on its own beyond the current scope, and further optimization is possible. All together, we show that any circuit with a limited number of gates to cut out can be simulated at the expense of a moderate overall overhead in sample complexity. As examples, we explicitly considered a Bell-state preparation circuit for two qubits increasingly far apart and variational ground-state solvers for the transverse-field Ising model on ring lattices. The former involves a single long-range gate, whereas the latter contains one such gate per variational layer.

Interestingly, the quasiprobability approach we use here is not the only route to gate simulation. In Ref. [46], similar in spirit to Ref. [26], quantum circuits are simulated via Monte Carlo simulation. However, instead of using the language of frames, a Hubbard-Stratonovich transformation is applied. In this context, the sign problem manifests itself in the form of a complex action that inhibits the efficient simulation of a large number of gates. A potential fruitful direction of future work would be to investigate the limits of this alternative gate simulation approach.

Importantly, our method requires platforms supporting midcircuit measurements and state preparations, which are readily provided by some quantum hardware companies such as, e.g., IBM and Honeywell [47,48]. This may pave the way to implement our method in a practical setting in the near future. However, the efficacy of our method will rely on the speed and accuracy of intermediate measurements. Although our numerical experiments for the CHSH violation indicate that our algorithm is insensitive to imperfect measurements, slow measurements may be more problematic since NISQ devices only have a limited coherence time.

Finally, we emphasize that our framework is not restricted to connectivity boosts only. It could also be applied to any gate that is too noisy for a given platform or combined with error-correcting codes to remove a gate that is particularly difficult to implement fault-tolerantly by the code. Another interesting application that will be studied elsewhere is circuit-depth boosts, where a deep circuit is simulated by shallower experimental circuits together with classical simulations of entire slices of the target circuit. In conclusion, our framework provides a versatile toolbox for both error-mitigation and circuit boosts well suited for noisy, intermediate-scale quantum hardware.

ACKNOWLEDGMENTS

The authors thank Ingo Roth for helpful insights. L.G. and L.A. acknowledge support from the Serrapilheira Institute (Grant No. Serra1709-17173) and the Brazilian agencies CNPq (PQ Grant No. 311416/2015-2 and INCT-IQ), FAPERJ (Grants No. PDR10 E-26/202.802/2016 and No. JCN E-26/202.701/2018), and FAPESP (Grants No. 2016/01343-7 and No. 2018/04208-9). J.F.C. acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Shared Hierarchical Academic Research Computing Network (SHARCNET), Compute Canada, the Google Quantum Research Award, and the CIFAR AI chair program. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute [49].

APPENDIX A: INTERFACES FOR HYBRID CLASSICAL-QUANTUM CIRCUITS WITH FRAMES

Here, we give a more formal presentation of the mathematical background of our algorithm in the language of frames.

1. Preliminaries

We consider an N-qubit system $\mathcal{S}$ of Hilbert space $\mathbb{H}_\mathcal{S}$ and denote the space of bounded, linear operators on $\mathbb{H}_\mathcal{S}$ by $\mathcal{L}(\mathbb{H}_\mathcal{S})$. We now consider the notion of a frame, which generalizes the notion of basis [22,23]. For our purposes, a frame $\mathcal{F}_\mathcal{S}$ for $\mathcal{L}(\mathbb{H}_\mathcal{S})$ is any set $\mathcal{F}_\mathcal{S} := \{M_a\}_a$ of Hermitian operators M_a that spans $\mathcal{L}(\mathbb{H}_\mathcal{S})$. Such a (in general, linearly dependent) spanning set is sometimes referred to as an overcomplete basis of $\mathcal{L}(\mathbb{H}_\mathcal{S})$. In turn, a frame $\mathcal{D}_\mathcal{S} := \{\tilde{M}_a\}_a$ such that

$$\mathcal{I} = \sum_a |\tilde{M}_a)(M_a|, \tag{A1}$$

where $\mathcal{I}$ is the identity map on $\mathcal{L}(\mathbb{H}_\mathcal{S})$, is called the dual to $\mathcal{F}_\mathcal{S}$ (and we then refer to $\mathcal{F}_\mathcal{S}$ as the primal to $\mathcal{D}_\mathcal{S}$). In Eq. (A1), the identity channel is written in the so-called Liouville or transfer matrix representation. That is, the round kets and bras denote 2^{2N}-dimensional column and row vectors, respectively, representing operators in $\mathcal{L}(\mathbb{H}_\mathcal{S})$ and their Hermitian adjoints. Accordingly, $(A|B)$ denotes the Hilbert-Schmidt inner product $\mathrm{Tr}[A^\dagger B]$ in $\mathcal{L}(\mathbb{H}_\mathcal{S})$. This is a popular notation in quantum information [22,23,50] that will be used here interchangeably with the (more usual) operator notation upon convenience.

We take throughout $M_{\boldsymbol{a}} \geqslant 0$ for all $\boldsymbol{a}$ and $\sum_{\boldsymbol{a}} M_{\boldsymbol{a}} = \mathbb{1}_{\mathcal{S}}$, with $\mathbb{1}_{\mathcal{S}}$ being the identity operator on $\mathbb{H}_{\mathcal{S}}$, so that $\mathcal{F}_{\mathcal{S}}$ is a positive operator-valued measure (POVM) on $\mathbb{H}_{\mathcal{S}}$. POVMs define generalized (i.e., beyond von Neumann) measurements [33,34]. This, together with Eq. (A1), allows us to express any density operator $\varrho \in \mathcal{L}(\mathbb{H}_{\mathcal{S}})$ as

$$|\varrho) = \sum_{\boldsymbol{a}} P_{\varrho}(\boldsymbol{a}) |\tilde{M}_{\boldsymbol{a}}), \tag{A2}$$

where $P_{\varrho}(\boldsymbol{a}) := (M_{\boldsymbol{a}}|\varrho)$ is the probability of measurement outcome $\boldsymbol{a}$ on ϱ. Equation (A2) is the basis of classical-shadow tomography, a powerful technique to get compact classical representations of states from measurements [21,35].

Note that $M_{\boldsymbol{a}} \geqslant 0$ for all $\boldsymbol{a}$ implies $\tilde{M}_{\boldsymbol{a}} \not\geqslant 0$ in general [22,23]. In addition, it will be useful to express the dual-frame elements as an affine combination of elements of $\mathcal{F}_{\mathcal{S}}$,

$$|\tilde{M}_{\boldsymbol{a}}) = \sum_{\boldsymbol{a}'} \mathfrak{T}_{\boldsymbol{a},\boldsymbol{a}'} |M_{\boldsymbol{a}'}), \quad \forall\, \boldsymbol{a}, \tag{A3}$$

for some adequately chosen $\mathfrak{T}$. With this parametrization, the primal- and dual-frame overlap matrices T and $\tilde{T}$, defined as $T_{\boldsymbol{a},\boldsymbol{a}'} := (M_{\boldsymbol{a}}|M_{\boldsymbol{a}'})$ and $\tilde{T}_{\boldsymbol{a},\boldsymbol{a}'} := (\tilde{M}_{\boldsymbol{a}}|\tilde{M}_{\boldsymbol{a}'})$, respectively, are related as $\tilde{T} = \mathfrak{T}\, T\, \mathfrak{T}$.

An experimentally convenient choice of $\mathcal{F}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{S}}$ is $M_{\boldsymbol{a}} = M_{a_1} \otimes \cdots \otimes M_{a_N}$ and $\tilde{M}_{\boldsymbol{a}} = \tilde{M}_{a_1} \otimes \cdots \otimes \tilde{M}_{a_N}$, for $\boldsymbol{a} := (a_1, \ldots, a_N)$. Here, M_{a_j} is the jth element of a single-qubit POVM frame, and $\tilde{M}_{a_j}$ is that of the corresponding dual frame. We refer to these as factorable frames. By virtue of Eqs. (A2) and (A3), these allow one to express any ϱ as an affine combination of product states $\sigma_{\boldsymbol{a}} := M_{\boldsymbol{a}}/t_{\boldsymbol{a}}$, where $t_{\boldsymbol{a}} := \mathrm{Tr}[M_{\boldsymbol{a}}]$ [20]. This fact has been used to reconstruct quantum states [20], processes [37], and overlaps [38] from single-qubit measurements. Additionally, this has been used to simulate quantum circuits [39] with generative machine learning models, where $\mathfrak{T}$ was taken as the canonical pseudoinverse of T. However, other choices of $\mathfrak{T}$ are possible. It can be seen (see Appendix C) that Eq. (A3) defines a dual to $\mathcal{F}_{\mathcal{S}}$ if and only if $\mathfrak{T}_{\boldsymbol{a},\boldsymbol{a}'} \in \mathbb{R}$, $\sum_{\boldsymbol{a}} \mathfrak{T}_{\boldsymbol{a},\boldsymbol{a}'} = 1$, and

$$T = T\, \mathfrak{T}\, T. \tag{A4}$$

In general, the elements of $\mathfrak{T}$ can be positive or negative. As shown below, the negativity of $\mathfrak{T}$ governs the sample complexity of Monte Carlo estimations of expectation values of observables. Finally, note also that if $\mathfrak{T}$ fulfills Eq. (A4), necessarily so does $\tilde{T} = \mathfrak{T}\, T\, \mathfrak{T}$ ($\tilde{T}$ and $\mathfrak{T}$ collapsing to each other for the canonical choice of $\mathfrak{T}$ being a pseudoinverse of T).

2. Interfaces for hybrid classical-quantum circuits

Our goal is to simulate quantum circuits using hybrid classical-quantum ones. More precisely, we are given an observable O, an N-qubit input state $\varrho_0 := |0\rangle\langle 0|$, and a target circuit $C := \{U_k\}_{k\in[f]}$, with $f \in \mathbb{N}$ single- or two-qubit unitary gates U_k. We denote by $s_k \subset \mathcal{S}$ the subset of qubits on which U_k acts and by $\boldsymbol{a}_{s_k}$ a corresponding substring of measurement outcomes on s_k. In addition, we use the shorthand notations $\bar{s}_k := \mathcal{S} \setminus s_k$ for the qubits on which U_k does not act and $\mathbb{1}_{\bar{s}_k}$ for the identity on $\mathbb{H}_{\bar{s}_k}$. From the f gates, $l < f$ are particularly experimentally demanding for NISQ implementations, and they are marked by the set of labels $L := \{k_1, k_2, \ldots, k_l\}$. The case we explicitly study below is that of two-qubit gates on qubits far apart in the connectivity graph in question. However, other relevant cases may be due to, e.g., error-mitigation convenience or other hardware-specific limitations. Either way, our goal is to estimate the expectation value $\mathrm{Tr}[\varrho_K\, O]$ of O on the output state $\varrho_K := U_f \cdots U_1\, \varrho_0\, U_1^{\dagger} \cdots U_f^{\dagger}$ by substituting every U_k with $k \in L$ by a classical simulation of it.

Our main tool to achieve this consists of interfaces between quantum objects and their (classical) frame representations. The first type of interface is based on Eq. (A2).

Definition 4: Quantum-classical interfaces. A QC interface on s_k refers to the assignment of a classical snapshot $\tilde{M}_{\boldsymbol{a}_{s_k}}$ to s_k according to the measurement outcome $\boldsymbol{a}_{s_k}$ of a factorable POVM frame $\mathcal{F}_{s_k}$ on a state $\varrho \in \mathbb{H}_{\mathcal{S}}$, occurring with probability $P_{\varrho}(\boldsymbol{a}_{s_k}) = (\mathbb{1}_{\bar{s}_k}|(M_{\boldsymbol{a}_{s_k}}|\varrho)$.

The second type of interface is the reverse interface, which simulates $\tilde{M}_{\boldsymbol{a}_{s_k}}$ as a linear combination of states $\sigma_{\boldsymbol{b}_{s_k}} := M_{\boldsymbol{b}_{s_k}}/t_{\boldsymbol{b}_{s_k}}$. This is done by importance-sampling $\boldsymbol{b}_{s_k}$ from $\tilde{T}^{(\mathcal{I}_{s_k})}$, given $\boldsymbol{a}_{s_k}$, with $\tilde{T}^{(\mathcal{I}_{s_k})}$ being the dual-frame overlap matrix on s_k. To see this, we apply on $|\tilde{M}_{\boldsymbol{a}_{s_k}})$ the Hermitian conjugate of Eq. (A1) and get $|\tilde{M}_{\boldsymbol{a}_{s_k}}) = \sum_{\boldsymbol{b}_{s_k}} \tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}\, t_{\boldsymbol{b}_{s_k}}\, |\sigma_{\boldsymbol{b}_{s_k}})$. Then, using a standard trick, we rewrite

$$\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} =: \left\|\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k}}\right\|_1 P_{\mathcal{I}_{s_k}}\left(\boldsymbol{b}_{s_k}\middle|\boldsymbol{a}_{s_k}\right) \mathrm{sgn}\left(\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}\right), \tag{A5}$$

where $\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k}}$ is a shorthand notation for the vector given by the $\boldsymbol{a}_{s_k}$th row of $\tilde{T}^{(\mathcal{I}_{s_k})}$, $\|\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k}}\|_1 := \sum_{\boldsymbol{b}_{s_k}} |\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}|$ is its l_1 norm, and $P_{\mathcal{I}_{s_k}}(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k}) := |\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}|/\|\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k}}\|_1$.

By construction, $P_{\mathcal{I}_{s_k}}(\circ|\boldsymbol{a}_{s_k})$ is a valid probability distribution, from which $\boldsymbol{b}_{s_k}$ can be sampled. This can be used to quantum-Monte-Carlo-simulate $\tilde{M}_{\boldsymbol{a}_{s_k}}$ [26].

Definition 5: Classical-quantum interface. A CQ interface on s_k refers to the repreparation of s_k in the state $\sigma_{\boldsymbol{b}_{s_k}}$, with probability $P_{\mathcal{I}_{s_k}}(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k})$, given a classical snapshot $\tilde{M}_{\boldsymbol{a}_{s_k}}$. Each sampled duple $(\boldsymbol{a}_{s_k}, \boldsymbol{b}_{s_k})$ is assigned the value $\|\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k}}\|_1\, t_{\boldsymbol{b}_{s_k}}\, \mathrm{sgn}(\tilde{T}^{(\mathcal{I}_{s_k})}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}})$.

The third and final ingredient integrates QC and CQ interfaces with a classical simulation of U_k. We denote by $\mathcal{U}_k$ the superoperator representing the action of the unitary U_k on $\mathcal{L}(\mathbb{H}_{\mathcal{S}})$. Multiplying $\mathcal{U}_k$ from the right by Eq. (A1) and from the left by the Hermitian conjugate of Eq. (A1), we get $\mathcal{U}_k = \sum_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} |M_{\boldsymbol{b}_{s_k}})\, \tilde{T}^{(\mathcal{U}_k)}_{\boldsymbol{b}_{s_k},\boldsymbol{a}_{s_k}}\, (M_{\boldsymbol{a}_{s_k}}|$, where $\tilde{T}^{(\mathcal{U}_k)}_{\boldsymbol{b}_{s_k},\boldsymbol{a}_{s_k}} := (\tilde{M}_{\boldsymbol{b}_{s_k}}|\,\mathcal{U}_k\,|\tilde{M}_{\boldsymbol{a}_{s_k}})$. With this, we get

$$\mathcal{U}_k|\varrho_{k-1}) = \sum_{\boldsymbol{a},\boldsymbol{a}'} \tilde{T}^{(\mathcal{U}_k)}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}\, t_{\boldsymbol{b}_{s_k}}\, |\sigma_{\boldsymbol{b}_{s_k}})\, (M_{\boldsymbol{a}_{s_k}}|\varrho_{k-1}), \tag{A6}$$

where $\varrho_{k-1} = U_{k-1} \cdots U_1\, \varrho_0\, U_1^{\dagger} \cdots U_{k-1}^{\dagger}$. That is, the action of $\mathcal{U}_k$ is absorbed into the repreparation by sampling from $\tilde{T}^{(\mathcal{U}_k)}$ instead of $\tilde{T}^{(\mathcal{I}_{s_k})}$. This leads to the following definition.

Definition 6: Quantum-classical-quantum interface. A QCQ interface for U_k on s_k refers to the measurement of $\mathcal{F}_{s_k}$, with outcome $\boldsymbol{a}_{s_k}$, followed by the repreparation of $\sigma_{\boldsymbol{b}_{s_k}}$ with probability $P_{\mathcal{U}_k}(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k}) := |\tilde{T}^{(\mathcal{U}_k)}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}|/\|\tilde{T}^{(\mathcal{U}_k)}_{\boldsymbol{a}_{s_k}}\|_1$. Each sampled duple $(\boldsymbol{a}_{s_k}, \boldsymbol{b}_{s_k})$ is assigned the value $v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} :=$

$\|\tilde{T}^{(\mathcal{U}_k)}_{\boldsymbol{a}_{s_k}}\|_1\, t_{\boldsymbol{b}_{s_k}}\, \mathrm{sgn}(\tilde{T}^{(\mathcal{U}_k)}_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}})$, and the corresponding interface realized in such an experimental run is thus mathematically represented by the operator $\mathcal{V}_k(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}) \coloneqq v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} |\sigma_{\boldsymbol{b}_{s_k}})(M_{\boldsymbol{a}_{s_k}}|$.

Our hybrid-circuit simulation then applies on ϱ_{k-1} the gate U_k if $k \notin L$, but it applies a QCQ interface for U_k instead if $k \in L$. Introducing the terminology

$$\mathcal{W}_k(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}) = \begin{cases} \mathcal{U}_k & \text{if } k \notin L \\ \mathcal{V}_k(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}) & \text{if } k \in L \end{cases} \tag{A7}$$

and using the fact that O is Hermitian, we can express the target expectation value $\mathrm{Tr}[\varrho_K\, O]$ as

$$(O|\varrho_K) = \sum_{\boldsymbol{\alpha}_{s_L}} \left(O \left| \prod_{k=1}^{f} \mathcal{W}_k(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}) \right| \varrho_0 \right), \tag{A8}$$

with the shorthand notation $\boldsymbol{\alpha}_{s_L} \coloneqq (\boldsymbol{a}_{s_{l'}}, \boldsymbol{b}_{s_{l'}}, \ldots, \boldsymbol{a}_{s_1}, \boldsymbol{b}_{s_1})$.

Equation (A8) can be experimentally estimated through an average $O^*_{M_s}$ over $M_s \in \mathbb{N}$ runs. M_s is chosen to guarantee that the statistical error and significance level (failure probability) of the estimation are given by the target values ε and δ, respectively. We refer to M_s as the *sample complexity* of the protocol, and its explicit value is given in Theorem 2 below.

The procedure is sketched by the pseudocode in Algorithm 2.

To quantify the runtime of the algorithm, we define the *interface negativity* of the gate $\mathcal{U}_k$ and the *total forward interface negativity* of the entire circuit C as

$$n_{U_k} \coloneqq \max_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}} \left\| \tilde{T}^{(\mathcal{U}_{s_k})}_{\boldsymbol{a}_{s_k}} \right\|_1 t_{\boldsymbol{b}_{s_k}} \quad \text{and} \quad n_{\to} \coloneqq \prod_{k\in L} n_{U_k}, \tag{A9}$$

respectively. This allows us to state the following theorem.

Theorem 2: Correctness and sample complexity. The finite-statistics average $O^*_{M_s}$ of Algorithm 2 is an unbiased estimator of $\mathrm{Tr}[\varrho_K\, O]$ (see Appendix D). Moreover, if

$$M \geqslant n_{\to}^2 \times \frac{2\,\|O\|^2\, \ln(2/\delta)}{\varepsilon^2}, \tag{A10}$$

with $\|O\|$ being the operator norm of O, then, with probability at least $1-\delta$, the statistical error of $O^*_{M_s}$ is at most ε.

Algorithm 2. Hybrid classical-quantum simulation with QCQ interfaces.

Input: ϱ_0, C, O, ε, δ
Output: $O^*_{M_s}$ such that $|O^*_{M_s} - \mathrm{Tr}[O\,\varrho_K]| \leqslant \varepsilon$ with probability at least $1-\delta$
Initialize $O^*_{M_s} = 0$, $v = 1$, and M_s as in Eq. (A10).
for $m \in (1,\ldots,M_s)$ **do**
  for $k \in (1,\ldots,f)$ **do**
    if $k \in L$ **them**
      Apply a QCQ interface for U_k on qubits s_k,
      obtaining the duple $(\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k})$;
      $v \leftarrow v \times v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$, with $v_{\boldsymbol{a}_{s_k},\boldsymbol{b}_{s_k}}$ as in Definition 6.
    else
      Apply the gate U_k on qubits s_k.
    end
  end
  Measure O, obtaining the measurement outcome (eigenvalue of O) o;
  $O^*_{M_s} \leftarrow O^*_{M_s} + o \times v$.
end
$O^*_{M_s} \leftarrow \frac{O^*_{M_s}}{M_s}$.

The proof follows straightforwardly from the Hoeffding bound. We note that the factor $\frac{2\,\|O\|^2\,\ln(2/\delta)}{\varepsilon^2}$ in Eq. (A10) is the equivalent sample-complexity bound one would obtain if $\mathrm{Tr}[\varrho_K\, O]$ was estimated from measurements on the actual state ϱ_K. Hence $n_{\to}^2$ quantifies the runtime overhead introduced by the interfaces. In that regard, the interface negativities play the same role in our hybrid classical-quantum simulation as the negativities of Ref. [26] in fully classical simulations with quasiprobability representations. An innovative and advantageous feature of Eq. (A9) is the presence of the POVM-element trace $t_{\boldsymbol{b}_{s_k}}$ in n_{U_k}, which comes from the state repreparation. Indeed, since $t_{\boldsymbol{b}_{s_k}} < 1$, the n_{U_k}'s (and therefore also $n_{\to}$) are significantly smaller than their counterparts for fully classical simulations [26]. This is consistent with the intuition that hybrid classical-quantum Monte Carlo simulations should cause lower sample-complexity increases than fully classical ones.

Either way, the most relevant property for our purposes is that $n_{\to}^2$ (and therefore also M_s) is independent not only of the numbers of gates f or qubits N but also, and most importantly, of the connectivity-graph distance between the qubits on which the interfaces act. In other words, for a fixed budget of measurement runs, simulating a gate U_k with a QCQ interface increases the statistical error at most by a constant factor n_{U_k}, regardless of how far apart in the circuit the qubits s_k are. In contrast, experimentally synthesizing U_k with noisy nearest-neighbor gates would give a systematic error due to infidelity accumulation that grows linearly with the distance between those qubits. Clearly, the drawback is that $n_{\to}^2$ grows exponentially with the number l of interfaces used. However, for many circuits, Algorithm 2 constitutes a better alternative than the bare NISQ implementation.

Finally, note that $n_{\to}^2$ is frame dependent. This is crucial to the efficiency of classical simulations [27–29]. For instance, in the quantum Monte Carlo method, it is known that the statistical overhead due to negative (quasi)probabilities can be ameliorated [32] or even removed [31] by local base changes. Something similar applies here: The interface negativities depend not only on the primal frame but also on the choice of dual to it.

APPENDIX B: INFORMATIONALLY COMPLETE POVMS

A positive operator-valued measure (POVM) is a set of operators $\{M_{\boldsymbol{a}}\}_{\boldsymbol{a}}$ with $M_{\boldsymbol{a}} \geqslant 0$ that satisfies the condition

$$\sum_{\boldsymbol{a}} M_{\boldsymbol{a}} = I. \tag{B1}$$

A POVM is informationally complete if $\{M_{\boldsymbol{a}}\}_{\boldsymbol{a}}$ spans $\mathcal{L}(\mathbb{H}_{\mathcal{S}})$. Let $\{M_{a_i}\}_{a_i}$ be a POVM that acts on a single-qubit Hilbert space. We can define a factorable POVM as a tensor product of a single-qubit POVM element as

$$M_{\boldsymbol{a}} = M_{a_1} \otimes \cdots \otimes M_{a_N}, \tag{B2}$$

for $\boldsymbol{a} := (a_1, \ldots, a_N)$. Clearly, if all M_{a_i} are informationally complete, then so is $M_{\boldsymbol{a}}$. An example of an informationally complete POVM is the Pauli-6 POVM, which is defined as

$$\{M_{\boldsymbol{a}}\}_{\boldsymbol{a}}^{\text{Pauli-6}} := \bigcup_{i=x,y,z} \left\{ \frac{1}{3}|\uparrow_i\rangle\langle\uparrow_i|, \frac{1}{3}|\downarrow_i\rangle\langle\downarrow_i| \right\}, \tag{B3}$$

where the vectors $|\uparrow_i\rangle, |\downarrow_i\rangle$ correspond to the eigenvectors of the Pauli operators with eigenvalue ± 1, respectively. We can implement this POVM by rotating to the Pauli basis with probability $1/3$. For the $\{X, Y, Z\}$ Pauli operators this means applying the gates $\{H, HS, I\}$, where H is a Hadamard gate, S is a Z-phase gate, and I is the identity gate. Measuring in the computational basis then produces outcomes $\boldsymbol{a}$ according to the Pauli-6 POVM.

APPENDIX C: DUAL-FRAME DECOMPOSITION

Here, we show that Eq. (A3) defines a dual frame with respect to $\mathcal{F}_S$ if Eq. (A4) holds. For the forward direction of this statement, we start with Eq. (A1) and plug in Eq. (A3) to obtain

$$\mathcal{I} = \sum_{\boldsymbol{a},\boldsymbol{b}} \mathfrak{T}_{\boldsymbol{a},\boldsymbol{b}} |M_{\boldsymbol{b}})(M_{\boldsymbol{a}}|. \tag{C1}$$

Applying $(M_{\boldsymbol{c}}|$ and $|M_{\boldsymbol{d}})$ to the left and right of Eq. (C1) then gives

$$(M_{\boldsymbol{c}}|M_{\boldsymbol{d}}) = \sum_{\boldsymbol{a},\boldsymbol{b}} \mathfrak{T}_{\boldsymbol{a},\boldsymbol{b}} (M_{\boldsymbol{c}}|M_{\boldsymbol{b}})(M_{\boldsymbol{a}}|M_{\boldsymbol{d}}); \tag{C2}$$

therefore we see that $T = T\mathfrak{T}T$ as required.

For the converse direction, we start with a map $\mathcal{J}$ on $\mathcal{L}(\mathbb{H}_S)$,

$$\mathcal{J} = \sum_{\boldsymbol{a},\boldsymbol{b}} \mathfrak{T}_{\boldsymbol{a},\boldsymbol{b}} |M_{\boldsymbol{b}})(M_{\boldsymbol{a}}|. \tag{C3}$$

Applying $(M_{\boldsymbol{c}}|$ and $|M_{\boldsymbol{d}})$ to the left and right of Eq. (C3) then gives

$$(M_{\boldsymbol{c}}|\mathcal{J}|M_{\boldsymbol{d}}) = \sum_{\boldsymbol{a},\boldsymbol{b}} \mathfrak{T}_{\boldsymbol{a},\boldsymbol{b}} (M_{\boldsymbol{c}}|M_{\boldsymbol{b}})(M_{\boldsymbol{a}}|M_{\boldsymbol{d}}). \tag{C4}$$

If we then plug in Eq. (A4), we find

$$(M_{\boldsymbol{c}}|\mathcal{J}|M_{\boldsymbol{d}}) = (M_{\boldsymbol{c}}|M_{\boldsymbol{d}}), \tag{C5}$$

from which we conclude that $\mathcal{J} \equiv \mathcal{I}$, i.e., $\mathcal{J}$ equals the identity map, and so Eq (A1) holds.

APPENDIX D: FINITE-STATISTICS ESTIMATOR

Let O be a generic observable we wish to measure, with support on an arbitrary subset of $\mathcal{S}$ and with arbitrary spectral norm $\|O\|_{\text{sp}} := o_{\max}$. Hence it admits a spectral decomposition as $|O) = \sum_\lambda o_\lambda |\lambda)$, where o_λ and $|\lambda)$ are its λth eigenvalue and eigenvector projector, respectively, with $|o_\lambda| \leqslant o_{\max}$ for all λ. Using Eq. (A8), we write the finite-statistics estimator of the expectation value $\langle O\rangle := \text{Tr}[O\,\varrho_{\text{f}}]$ of O as

$$O_M^* := \frac{1}{M} \sum_{i=1}^{M} o_{\lambda^{(i)},\boldsymbol{\alpha}_{s_L}^{(i)}} \prod_{k\in L} v_{\boldsymbol{a}_{s_k}^{(i)},\boldsymbol{b}_{s_k}^{(i)}}, \tag{D1}$$

where $o_{\lambda^{(i)},\boldsymbol{\alpha}_{s_L}^{(i)}}$ is the eigenvalue obtained from the single-shot i obtained from a state that is measured and reprepared according to $\boldsymbol{\alpha}_{s_L}^{(i)}$. The probability of observing $o_{\lambda^{(i)},\boldsymbol{\alpha}_{s_L}^{(i)}}$ is given by

$$P\big(o_{\lambda^{(i)},\boldsymbol{\alpha}_{s_L}^{(i)}}\big) = \left(\lambda^{(i)} \left| \prod_{k=1}^{f} \mathcal{W}_k\big(\boldsymbol{a}_{s_k}^{(i)}, \boldsymbol{b}_{s_k}^{(i)}\big) \right| \varrho_0 \right), \tag{D2}$$

with $\boldsymbol{a}_{s_k}^{(i)} \sim P_{\varrho_{k-1}}(\boldsymbol{a}_{s_k})$ and $\boldsymbol{b}_{s_k}^{(i)} \sim P_{\mathcal{U}_k}(\boldsymbol{b}_{s_k}|\boldsymbol{a}_{s_k})$, where

$$|\varrho_{k-1}) = \prod_{k=1}^{l-1} \mathcal{W}_l\big(\boldsymbol{a}_{s_l}^{(i)}, \boldsymbol{b}_{s_l}^{(i)}\big)|\varrho_0). \tag{D3}$$

Importantly, O_M^* is an unbiased estimator.

APPENDIX E: THE CLAUSER-HORNE-SHIMONY-HOLT INEQUALITIES

The CHSH inequalities constrain a set of four correlators in an experiment of Alice (A) and Bob (B) type and provide a condition to check whether the correlations between the observations of Alice and Bob can be explained by a local theory or necessitate a nonlocal theory such as quantum mechanics [51]. Consider the quantity

$$S(A,B) = C^{00}(A,B) + C^{01}(A,B) + C^{10}(A,B) \tag{E1}$$

$$- C^{11}(A,B), \tag{E2}$$

where

$$C^{00}(A,B) = \frac{1}{\sqrt{2}}(-\langle Z_A Z_B\rangle - \langle Z_A X_B\rangle), \tag{E3}$$

$$C^{01}(A,B) = \frac{1}{\sqrt{2}}(-\langle X_A Z_B\rangle - \langle X_A X_B\rangle), \tag{E4}$$

$$C^{10}(A,B) = \frac{1}{\sqrt{2}}(\langle Z_A Z_B\rangle - \langle Z_A X_B\rangle), \tag{E5}$$

$$C^{11}(A,B) = \frac{1}{\sqrt{2}}(\langle X_A Z_B\rangle - \langle Z_A X_B\rangle) \tag{E6}$$

are the correlations obtained from the state shared by Alice and Bob. The observables X and Z are the Pauli matrices. We call $S(A,B)$ the Bell polynomial. The CHSH inequality is given by $S(A,B) \leqslant 2$, which, if satisfied, implies that a local hidden variable theory can explain the observed correlations. On the other hand, for $S(A,B) > 2$ we have to invoke quantum theory to explain the correlations. The maximum value of $S(A,B)$ is $2\sqrt{2}$, which is obtained for a maximally entangled two-qubit state.

APPENDIX F: LOCALLY PURIFIED DENSITY OPERATORS

Numerical simulations with a full density matrix of size $2^N \times 2^N$ quickly become prohibitive due to the large memory requirements. Hence we have to resort to tensor networks to find efficient representations of mixed quantum states. The canonical choice for representing operators with tensor networks is the matrix product operator (MPO) [52]. A drawback of this approach is that applying completely positive maps to the state can still lead to the MPO becoming nonpositive due to truncation errors. The locally purified density operator

(LPDO) tensor network solves this issue by representing the state as $\varrho = \chi\chi^\dagger$, where the purification operator χ is given by a tensor network

$$[\chi]^{p_1,\ldots,p_N}_{\kappa_1,\ldots,\kappa_N} = \sum_{b_1,\ldots,b_{N-1}} A^{[1]p_1,\kappa_1}_{b_1} A^{[2]p_2,\kappa_2}_{b_1,b_2} \cdots A^{[N]p_N\kappa_N}_{b_{N-1}}, \tag{F1}$$

with $1 \leqslant p_l \leqslant P$, $1 \leqslant \kappa_l \leqslant \kappa$, and $1 \leqslant b_l \leqslant D$ [41]. Here, P is called the physical dimension, κ is the Kraus dimension, and D is the bond dimension.

Analogous to the bond dimension truncation in MPOs, truncating the Kraus dimension after applying a channel leads to errors in our state representation that can affect the accuracy of numerical simulations. However, we can control the accuracy of the simulation by increasing D and κ and keeping track of a runtime lower bound estimate of the state fidelity. Let $\varrho = \chi^\dagger\chi$ and $\sigma = \eta^\dagger\eta$; then the fidelity is given by

$$F(\varrho,\sigma) = \mathrm{Tr}\sqrt{\sqrt{\sigma}\varrho\sqrt{\sigma}}. \tag{F2}$$

From Lemma 1 in Ref. [41] we know that

$$F(\varrho,\sigma) \geqslant \tfrac{1}{2}\big(2 - \|\chi - \eta\|_2^2\big). \tag{F3}$$

Let χ be a locally purified description of a quantum state with local tensors $\{A^{[N]}\}$ that is in mixed canonical form with respect to a local tensor $A^{[l_{cp}]}$. If a single tensor $A^{[l]}$ is compressed by discarding singular values in either the Kraus or bond dimensions, then by Lemma 6 of Ref. [41] we know that

$$\delta := \left(\sum_{i,\mathrm{discarded}} s_i^2\right)^{\frac{1}{2}}, \tag{F4}$$

and subsequently

$$\|\chi - \chi'\|_2^2 = 2(1 - \sqrt{1-\delta^2}), \tag{F5}$$

where χ' is the compressed tensor. By the triangle inequality, the two norm errors introduced by the discarded weights can at most sum up. Hence the true operator norm is lower bounded by the sum of all discarded weight errors

$$\|\varrho_{\mathrm{exact}} - \varrho_{\mathrm{truncated}}\|_2 \leqslant \sum_d \sqrt{2\big(1 - \sqrt{1-\delta_d^2}\big)}, \tag{F6}$$

with d being the number of truncations and δ_k being the discarded weights. This brings the final runtime fidelity estimate to

$$F(\varrho,\sigma) \geqslant \frac{1}{2}\big(2 - \|\chi - \eta\|_2^2\big) \tag{F7}$$

$$\geqslant \frac{1}{2}\left(2 - \left(\sum_d \sqrt{2\big(1 - \sqrt{1-\delta_d^2}\big)}\right)^2\right). \tag{F8}$$

In all our experiments, we apply depolarizing channels to both qubits only after applying a two-qubit gate, since single-qubit gate noise tends to be small in experimental settings. The single-qubit depolarizing channel is given by

$$\varrho = \sum_{m=1}^{M} \kappa_m \varrho K_m^\dagger, \tag{F9}$$

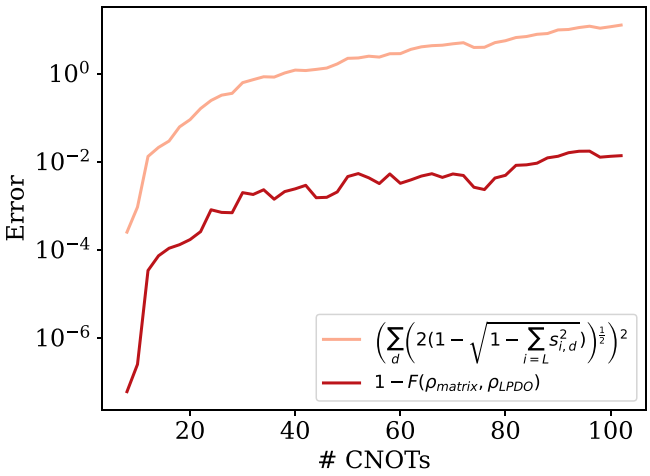

FIG. 7. Illustration of the lower bound of Eq. (F8). The circuit consists of an initial state $|+\rangle^{\otimes 4}$ to which we apply a varying number of CNOT gates with random control and target qubits. We set the noise to $\lambda = 0.005$ and take $D = 4$ and $\kappa = 16$. The red line indicates the true accuracy of the LPDO simulation by comparing it with the exact full density matrix simulation. The orange line gives the runtime fidelity estimate. We see that the accuracy of the simulation degrades as we add more two-qubit gates and depolarizing channels. The runtime fidelity gives an estimate two orders of magnitude above the exact error, indicating that for this example, the bound is a conservative estimate of the simulation error.

where $\{K_m\}$ is a set of Kraus operators with

$$K_1 = \sqrt{\frac{(4-3\lambda)}{4}}\mathbb{1}, \quad K_2 = \sqrt{\frac{\lambda}{4}}X, \tag{F10}$$

$$K_3 = \sqrt{\frac{\lambda}{4}}Y, \quad K_4 = \sqrt{\frac{\lambda}{4}}Z. \tag{F11}$$

Here, $\{X, Y, Z\}$ are the Pauli matrices, and $\mathbb{1}$ is the identity. The scalar $\lambda \in [0, 1]$ controls the strength of the depolarization. With these channels, illustrate the bound of Eq. (F8) by comparing the final state overlap of an exact full density matrix simulation and a LPDO simulation for a random four-qubit circuit with a varying number of CNOT gates. In Fig. 7, we see that the runtime estimate of the fidelity is about two orders of magnitude above the true fidelity.

APPENDIX G: RANDOM WALK METROPOLIS-HASTINGS ALGORITHM FOR NEGATIVITY MINIMIZATION

In this Appendix, we present a method to minimize the sample-complexity overhead by the interface of a unitary gate U exploiting the freedom in the choice of dual POVM, namely, the choice of $\mathfrak{T}$ subject to Eq. (A4). For concreteness, we focus on the case where all POVM elements have the same trace, so $\mathrm{Tr}[M_{\boldsymbol{b}}] = 1/D$ for all $\boldsymbol{b}$, with D being the number of POVM elements. Moreover, we optimize a modified version of the interface negativity n_U where, instead of maximizing $\|\tilde{T}_{\boldsymbol{a}}^U\|_1$ over $\boldsymbol{a}$ [as in Eq. (A9)], we average $\|\tilde{T}_{\boldsymbol{a}}^U\|_1^2$ over $\boldsymbol{a}$. Such an average is the sample-complexity overhead directly given by the Hoeffding bound for when the sampled random variables can lie within segments of different lengths. The reason for this modification is that, while in Theorem 2 we are

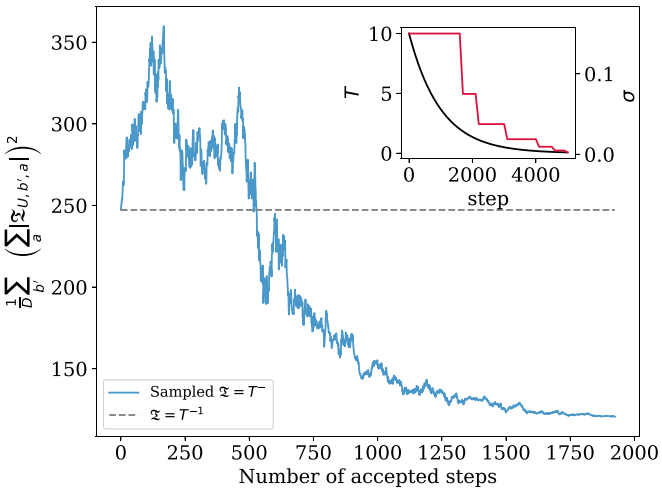

FIG. 8. Monte Carlo random walk for interface negativity optimization of the ZZ gate used in Sec. IV B of the main text. The total number of steps in the annealing schedule is 5000. The gray dashed line indicates the mean average squared negativity of the pseudoinverse, whereas the blue line indicates the one for the newly accepted $\mathfrak{T}$'s during the Monte Carlo random walk. The inset shows the adaptive scheme that fine-tunes the search with the temperature and variance, given in black and red, respectively.

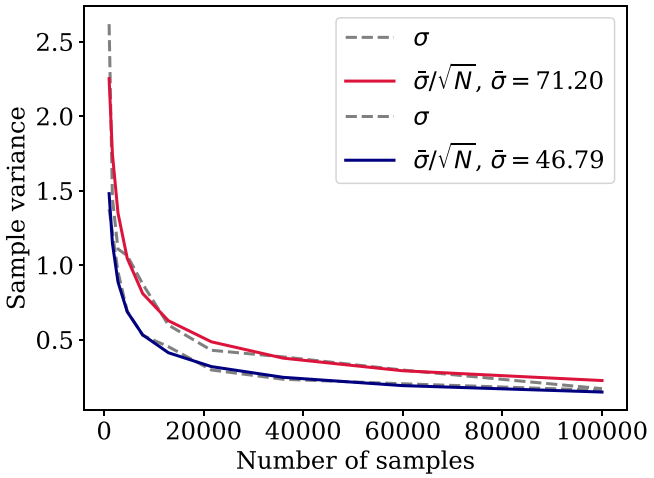

FIG. 9. Improvement of energy-estimator variance for the eight-qubit TFIM circuit experiment of Fig. 4(b) in the main text. Sample variance is estimated over 50 runs. The red line shows the sample variance corresponding to the canonical dual frame given by the pseudoinverse of T. In blue we see the variance of the energy corresponding to the dual frame obtained from the Monte Carlo search.

interested in the worst-case complexity, here we are interested in the more practical problem of the average case.

For optimizing $\tilde{T}^U$ over $\mathfrak{T}$, we express it as $\tilde{T}^U = \mathfrak{T}_1 T^U \mathfrak{T}_2$, with T^U given by $T^{U_k} := \mathrm{Tr}[U_k M_{\boldsymbol{a}_{s_k}} U_k^\dagger M_{\boldsymbol{b}_{s_k}}]$. Note that by not enforcing that $\mathfrak{T}_1 = \mathfrak{T}_2$, we are explicitly allowing for the more general case of possibly different input and output dual POVMs. Hence we wish to solve the constrained nonconvex optimization

$$\min_{\mathfrak{T}} \quad \frac{1}{D}\sum_{\boldsymbol{a}} \|(\mathfrak{T}_1 T^U \mathfrak{T}_2)_{\boldsymbol{a}}\|_1^2, \tag{G1}$$

$$\text{such that} \quad T = T\mathfrak{T}_i T, \quad \text{for } i = 1, 2, \tag{G2}$$

where $(\mathfrak{T}_1 T^U \mathfrak{T}_2)_{\boldsymbol{a}}$ is a shorthand notation for the $\boldsymbol{a}$th row of $\mathfrak{T}_1 T^U \mathfrak{T}_2$ and $\|(\mathfrak{T}_1 T^U \mathfrak{T}_2)_{\boldsymbol{a}}\|_1$ is its l_1 norm. Equation (G2) is a necessary but not sufficient condition for $\mathfrak{T}_i$ to be the Penrose-Moore pseudoinverse of T. Indeed, such a condition implies that $\mathfrak{T}_i$ is a so-called generalized inverse of T [36,53]. So, the first question we need to consider is how to variationally explore the space of generalized inverses of T in a practical way.

Fortunately, this question has been previously studied. In particular, in Ref. [30] it was shown that for an arbitrary matrix $A \in \mathbb{R}^{m\times n}$ and given any particular generalized inverse A^- of it, every generalized inverse B^- can be obtained from some $C \in \mathbb{R}^{m\times n}$ by the map

$$B^-(C) := A^- + C - A^- A C A A^-. \tag{G3}$$

That is, the entire space of generalized inverses is parametrized by C. This leads us to a practical way to obtain a random walk across the space of generalized inverses: In the first iteration, take the Penrose-Moore pseudoinverse A^{-1} as the starting generalized inverse and a randomly sampled C. This produces the first B^-. As inputs for the second iteration, use the first iteration's output B^- as the generalized inverse and a fresh, independently sampled C. This produces a new B^-. Then continue to iterate.

Using this recipe for $A = T$ and $A^{-1} = T^{-1}$, we can ergodically explore the space of generalized inverses $\mathfrak{T}_i$ of T. In turn, the resulting random walk can be used as Markov chain Monte Carlo (MCMC) dynamics for a simulated-annealing optimization [54,55] that approximates a solution to Eq. (G1). More precisely, for each random walk iteration, we (probabilistically) accept or reject the newly produced $\mathfrak{T}_i$ via a standard Metropolis-Hastings algorithm with $\frac{1}{D}\sum_{\boldsymbol{a}} \|(\mathfrak{T}_1 T^U \mathfrak{T}_2)_{\boldsymbol{a}}\|_1^2$ as the energy function.

For a two-qubit gate U and the Pauli-6 POVM, each dual-overlap matrix can be expressed as $\mathfrak{T}_i = \mathfrak{T}_i^{(1)} \otimes \mathfrak{T}_i^{(2)}$, where $\mathfrak{T}_i^{(1)}$ and $\mathfrak{T}_i^{(2)}$ are the 6×6 real dual-overlap matrices of the two qubits on which U acts. We can independently sample all four matrices, $\mathfrak{T}_1^{(1)}$, $\mathfrak{T}_1^{(2)}$, $\mathfrak{T}_2^{(1)}$, and $\mathfrak{T}_2^{(2)}$. Hence the search-space dimension is $4 \times 6 \times 6 = 142$.

For the simulated-annealing schedule, we take random matrices $C \sim \mathcal{N}(0, \sigma^2)^{6\times 6}$. We set the initial temperature to be $T = 10$ and decrease it by a factor of 0.999 at each Monte Carlo step (MCS). In addition to the temperature, the Monte dynamics are controlled by the variance σ^2 of the normal distribution $\mathcal{N}(0, \sigma^2)^{6\times 6}$ for C. We start with a large initial $\sigma^2 = 0.1$ to coarsely explore the search space. However, as the temperature decreases, we want to refine the search without freezing the Monte Carlo dynamics. Therefore we use an adaptive scheme where σ^2 is decreased according to the acceptance ratio. Specifically, we halve the value of σ if the acceptance ratio per 100 MCSs is smaller than 0.23, a well-known heuristic for continuous-variable MCMC [56]. The search is terminated if the negativity decreases less than 10^{-2} after 100 accepted steps.

As a result, we consistently find dual frames whose averaged squared negativities are about half the value of the canonical dual frame from the pseudoinverse (see Fig. 8). This is also observed to greatly improve the sample complexity in practice (see Fig. 9).

[1] A. Montanaro, Quantum algorithms: an overview, npj Quantum Inf. **2**, 15023 (2016).

[2] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).

[3] A. Bouland, W. van Dam, H. Joorati, I. Kerenidis, and A. Prakash, Prospects and challenges of quantum finance, arXiv:2011.06492.

[4] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[5] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, Parameterized quantum circuits as machine learning models, Quantum Sci. Technol. **4**, 043001 (2019).

[6] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum (NISQ) algorithms, Rev. Mod. Phys. **94**, 015004 (2022).

[7] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[8] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[9] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[10] K. Temme, S. Bravyi, and J. M. Gambetta, Error Mitigation for Short-Depth Quantum Circuits, Phys. Rev. Lett. **119**, 180509 (2017).

[11] S. Endo, S. C. Benjamin, and Y. Li, Practical Quantum Error Mitigation for Near-Future Applications, Phys. Rev. X **8**, 031027 (2018).

[12] Y. Li and S. C. Benjamin, Efficient Variational Quantum Simulator Incorporating Active Error Minimization, Phys. Rev. X **7**, 021050 (2017).

[13] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta, Error mitigation extends the computational reach of a noisy quantum processor, Nature (London) **567**, 491 (2019).

[14] S. Bravyi, G. Smith, and J. A. Smolin, Trading Classical and Quantum Computational Resources, Phys. Rev. X **6**, 021043 (2016).

[15] V. Dunjko, Y. Ge, and J. I. Cirac, Computational Speedups Using Small Quantum Devices, Phys. Rev. Lett. **121**, 250501 (2018).

[16] T. Peng, A. W. Harrow, M. Ozols, and X. Wu, Simulating Large Quantum Circuits on a Small Quantum Computer, Phys. Rev. Lett. **125**, 150504 (2020).

[17] W. Tang, T. Tomesh, M. Suchara, J. Larson, and M. Martonosi, CutQC: Using small quantum computers for large quantum circuit evaluations, in *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2021)* (Association for Computing Machinery, New York, 2021), pp. 473–486.

[18] M. A. Perlin, Z. H. Saleem, M. Suchara, and J. C. Osborn, Quantum circuit cutting with maximum-likelihood tomography, npj Quantum Inf. **7**, 64 (2021).

[19] K. Mitarai and K. Fujii, Constructing a virtual two-qubit gate by sampling single-qubit operations, New J. Phys. **23**, 023021 (2021).

[20] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Reconstructing quantum states with generative models, Nat. Mach. Intell. **1**, 155 (2019).

[21] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[22] C. Ferrie and J. Emerson, Frame representations of quantum mechanics and the necessity of negativity in quasi-probability representations, J. Phys. A: Math. Theor. **41**, 352001 (2008).

[23] C. Ferrie, Quasi-probability representations of quantum theory with applications to quantum information science, Rep. Prog. Phys. **74**, 116001 (2011).

[24] A. Mari and J. Eisert, Positive Wigner Functions Render Classical Simulation of Quantum Computation Efficient, Phys. Rev. Lett. **109**, 230503 (2012).

[25] V. Veitch, N. Wiebe, C. Ferrie, and J. Emerson, Efficient simulation scheme for a class of quantum optics experiments with non-negative Wigner representation, New J. Phys. **15**, 013037 (2013).

[26] H. Pashayan, J. J. Wallman, and S. D. Bartlett, Estimating Outcome Probabilities of Quantum Circuits Using Quasiprobabilities, Phys. Rev. Lett. **115**, 070501 (2015).

[27] N. Hatano and M. Suzuki, Representation basis in quantum Monte Carlo calculations and the negative-sign problem, Phys. Lett. A **163**, 246 (1992).

[28] E. Y. Loh, J. E. Gubernatis, R. T. Scalettar, S. R. White, D. J. Scalapino, and R. L. Sugar, Sign problem in the numerical simulation of many-electron systems, Phys. Rev. B **41**, 9301 (1990).

[29] M. Troyer and U.-J. Wiese, Computational Complexity and Fundamental Limitations to Fermionic Quantum Monte Carlo Simulations, Phys. Rev. Lett. **94**, 170201 (2005).

[30] C. Radhakrishna Rao, *Calculus of Generalized Inverses of Matrices Part I: General Theory*, Sankhya: The Indian Journal of Statistics, Series A (1961-2002) (Indian Statistical Institute, Kolkata, India, 1967), Vol. 29, pp. 317–342.

[31] M. Marvian, D. A. Lidar, and I. Hen, On the computational complexity of curing non-stoquastic Hamiltonians, Nat. Commun. **10**, 1571 (2019).

[32] D. Hangleiter, I. Roth, D. Nagaj, and J. Eisert, Easing the Monte Carlo sign problem, Sci. Adv. **6**, eabb8341 (2020).

[33] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 10th ed. (Cambridge University Press, Cambridge, 2011).

[34] A. Peres, *Quantum Theory: Concepts and Methods*, Fundamental Theories of Physics Vol. 72 (Kluwer Academic, Dordrecht, 1993).

[35] S. Aaronson, Shadow tomography of quantum states, in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2018)* (Association for Computing Machinery, New York, 2018), pp. 325–338.

[36] R. Penrose, A generalized inverse for matrices, Math. Proc. Cambridge Philos. Soc. **51**, 406 (1955).

[37] G. Torlai, C. J. Wood, A. Acharya, G. Carleo, J. Carrasquilla, and L. Aolita, Quantum process tomography with unsupervised learning and tensor networks, arXiv:2006.02424.

[38] J. F. C. Leonardo Guerini, R. Wiersema and L. Aolita, Quasiprobabilistic state-overlap estimator for NISQ devices, arXiv:2112.11618.

[39] J. Carrasquilla, D. Luo, F. Pérez, A. Milsted, B. K. Clark, M. Volkovs, and L. Aolita, Probabilistic simulation of quantum circuits using a deep-learning architecture, Phys. Rev. A **104**, 032610 (2021).

[40] C. Piveteau and D. Sutter, Circuit knitting with classical communication, arXiv:2205.00016.

[41] A. H. Werner, D. Jaschke, P. Silvi, M. Kliesch, T. Calarco, J. Eisert, and S. Montangero, Positive Tensor Network Approach for Simulating Open Quantum Many-Body Systems, Phys. Rev. Lett. **116**, 237201 (2016).

[42] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler *et al.*, Quantum supremacy using a programmable superconducting processor, Nature (London) **574**, 505 (2019).

[43] W. W. Ho and T. H. Hsieh, Efficient variational simulation of non-trivial quantum states, SciPost Phys. **6**, 029 (2019).

[44] D. Wierichs, C. Gogolin, and M. Kastoryano, Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer, Phys. Rev. Res. **2**, 043246 (2020).

[45] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, Exploring Entanglement and Optimization within the Hamiltonian Variational Ansatz, PRX Quantum **1**, 020319 (2020).

[46] N. J. Cerf and S. E. Koonin, Monte Carlo simulation of quantum computation, Math. Comput. Simul. **47**, 143 (1998).

[47] Honeywell Quantum, Mid-circuit measurements on the System Model H1, 2021, https://www.honeywell.com/us/en/company/quantum/quantum-computer/.

[48] IBM Quantum, Mid-circuit Measurements Tutorial, 2021, https://quantum-computing.ibm.com/lab/docs/iql/manage/systems/midcircuit-measurement/.

[49] https://vectorinstitute.ai/#partners.

[50] C. M. Caves, Quantum error correction and reversible operations, J. Supercond. **12**, 707 (1999).

[51] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Proposed Experiment to Test Local Hidden-Variable Theories, Phys. Rev. Lett. **23**, 880 (1969).

[52] F. Verstraete, J. J. García-Ripoll, and J. I. Cirac, Matrix Product Density Operators: Simulation of Finite-Temperature and Dissipative Systems, Phys. Rev. Lett. **93**, 207204 (2004).

[53] E. Moore, On the reciprocal of the general algebraic matrix, Bull. Am. Math. Soc. **26**, 394 (1920).

[54] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. **21**, 1087 (1953).

[55] C. Sherlock, P. Fearnhead, and G. O. Roberts, The random walk Metropolis: Linking theory and practice through a case study, Stat. Sci. **25**, 172 (2010).

[56] A. Gelman, W. R. Gilks, and G. O. Roberts, Weak convergence and optimal scaling of random walk Metropolis algorithms, Ann. Appl. Probab. **7**, 110 (1997).