Learning Correlated Noise in a 39-Qubit Quantum Processor

Robin Harper[1,*] and Steven T. Flammia[2,3]

[1] *Centre for Engineered Quantum Systems, School of Physics, University of Sydney, Sydney, New South Wales 2006, Australia*

[2] *AWS Center for Quantum Computing, Pasadena, California 91125, USA*

[3] *Institute for Quantum Information and Matter (IQIM), California Institute of Technology, Pasadena, California 91125, USA*

(Received 18 April 2023; accepted 29 August 2023; published 17 October 2023)

Building error-corrected quantum computers relies crucially on measuring and modeling noise on candidate devices. In particular, optimal error correction requires knowing the noise that occurs in the device as it executes the circuits required for error correction. As devices increase in size, we will become more reliant on efficient models of this noise. However, such models must still retain the information required to optimize the algorithms used for error correction. Here, we propose a method of extracting detailed information of the noise in a device running syndrome extraction circuits. We introduce and execute an experiment on a superconducting device using 39 of its qubits in a surface code doing repeated rounds of syndrome extraction but omitting the midcircuit measurement and reset. We show how to extract from the 20 data qubits the information needed to build noise models of various sophistication in the form of graphical models. These models give efficient descriptions of noise in large-scale devices and are designed to illuminate the effectiveness of error correction against correlated noise. Our estimates are furthermore precise: we learn a consistent global distribution where all one- and two-qubit error rates are known to a relative error of 0.1%. By extrapolating our experimentally learned noise models toward lower error rates, we demonstrate that accurate correlated noise models are increasingly important for successfully predicting subthreshold behavior in quantum error-correction experiments.

DOI: 10.1103/PRXQuantum.4.040311

I. INTRODUCTION

In order to fully realize the potential of quantum devices, one must execute many highly accurate quantum gates [1–5]. Current multiqubit devices have single-qubit gates with average error rates around 10^{-3} [6–9]; this is orders of magnitude too high to directly execute the number of operations required for computations such as integer factoring [5]. Fault-tolerant quantum error correction overcomes this by trading more physical qubits for increased logical qubit fidelity [1,10–12]. Although there are many proposed protocols for error correction and many recent error-correction experiments for various small codes [13–17], the most widely studied proposals are based on the surface code and variants thereof [18–22]. Recent work has shown a decreasing logical error rate in a variant of the surface code [22] as the distance of the code increases [23].

Error correction can be dramatically improved provided that one knows certain details of the noise afflicting the device [22–28] and one tailors the code and decoder to this noise [22,29–32]. However, most existing noise-characterization techniques fall into two categories: tomographic reconstruction or strong noise averaging (e.g., randomized benchmarking). Those based on full tomographic noise reconstruction [33–36] struggle to scale past a few qubits. Techniques that strongly average the noise [37–47] attempt to gain scalability but often fail to provide enough of a nuanced picture of the noise to allow diagnostic conclusions or to suggest improvements to the device. We review these methods in Appendix A.

Fortunately, there are techniques that reconstruct efficient descriptions of detailed Pauli noise models for large-scale systems [48,49]. These methods learn the noise in the form of a *graphical model*, which is a model that is flexible enough to provide faithful descriptions of correlated noise while retaining the desirable properties of being efficient to learn and allowing the tailoring of codes and decoders to relevant features of the noise.

*Corresponding author: robin.harper@sydney.edu.au

An example of such a graphical model is an Ising model, where a probability distribution with 2^n outcomes—the equilibrium Gibbs distribution—is specified by a Hamiltonian and a temperature. The individual probabilities of such a model cannot be calculated without knowing the partition function (which is not generally efficient to compute). However, ratios of probabilities and certain conditional probabilities can be calculated efficiently and they often enjoy efficient sampling algorithms (using, e.g., Metropolis sampling). For quantum noise that is described by such a graphical model, these features are sufficient to enable detailed modeling and prediction, as well as optimizations such as the creation of minimum weight perfect matching decoders or tensor network decoders that are tailored to the underlying noise correlations. For instance, Ref. [50] explores the duality between such graphical models and tensor networks. Accordingly, such graphical models can directly instantiate tensor networks, allowing utilization of the contraction methods currently proposed as a means of decoding syndrome measurements [27,28,51,52].

Here, we experimentally demonstrate learning a comprehensive description of the Pauli noise in a 39-qubit superconducting device running the circuits required to implement surface-code quantum error correction but without midcircuit ancilla measurement and reset. The noise is learned in the form of a graphical model, meaning this description is efficient and contains a globally consistent description of the errors in the device, including larger-scale correlations.

Our reconstructed noise estimates are highly accurate. By way of example, a bootstrap analysis (at the 2σ level) shows a maximum *relative* error of $\pm 0.1\%$ on both single-qubit error and two-qubit error rates. We fit several graphical models having increasing expressive power to these data to see how correlations affect quantum error correction. We measure the logical failure rate assuming a code capacity model (i.e., no measurement and reset errors) and utilizing an approximate maximum-likelihood decoder (a tensor network decoder [53] with bond dimension 8). Even in the absence of measurement and reset error, the circuit noise translated to a logical error rate of at least 0.166 ± 0.004, which is worse than the observed average single-qubit error rate of 0.136 ± 0.001.

Our reconstructed noise models allow us to address previously difficult-to-explore questions of interest for error correction in this system. For example, given the observed correlations, how much lower should the single-qubit error rates be before the logical failure rate is below the single-qubit bare error rate? To address this, we parametrize our observed noise as being generated by a continuous-time evolution evolving for a finite time T; this allows us to retain the observed correlation structure in a way that extrapolates smoothly to zero noise as $T \to 0$.

Perhaps surprisingly, the simplest models (which ignore correlated errors) through to the most complex models have all given approximately the same pseudothreshold (the physical error rate equal to the logical error rate of approximately 0.1). However, simpler models underestimate the logical error rate compared with the global distribution and give widely diverging predictions of the logical failure rate as the physical error rate extrapolates toward zero. For instance, with an average physical error rate of 0.031, the simpler models have predicted logical error rates of 0.006 ± 0.001, whereas the logical error rate from the extrapolated global probability distribution, has been over twice as much (0.0121 ± 0.002). In this regime, models that capture correlated errors, such as an Ising model, have given logical error rates of $0.014 - 0.018$ that are commensurate with or higher than the global distribution.

Our conclusion is that models capable of capturing correlated errors are indispensable in accurately estimating the expected logical failure rate of error-correction protocols executed on this device. Our analysis shows that once physical error rates are low enough such that error correction might be possible, models that fail to take into account correlated errors (such as those caused by crosstalk) can potentially underestimate expected logical failure rates by a significant fraction. However, models that can capture such crosstalk provide more reliable estimates of the performance of the error-correcting circuitry as well as providing the more nuanced information required to write custom decoders.

II. SURFACE-CODE PROTOCOL

The surface code is an error-correcting code that appears to be particularly well suited for two-dimensional grids of superconducting qubits. Figure 1 shows a typical surface-code layout and how the abstract code can be mapped to the gridlike qubit devices used by Google (for more details of the device, see Appendix B). In this paper, we consider the standard (Calderbank-Shor-Steane, CSS) surface code with X- and Z-type stabilizer generators, although we note that more recent experiments on Google devices [23] used the closely related $XZZX$ code instead [22]. Our results could equally be applied in this case.

Error correction is a technique with many moving parts, all of which have the potential of introducing noise into the system. Like any complex system, if all the parts are in action, it can be difficult to diagnose the source of problems. For this reason, as well as technical limitations of the device, we focus on a simplified analogue of error-correction circuits. Specifically, we run the circuits required for nondemolition four-body stabilizer measurements of the data qubits but without actually performing the ancilla measurements or the resets required in a real error-correction experiment. We call these circuits *stabilizer preparation*, since we only prepare to execute, but do not actually execute, the measurement. Here, we show how to use efficient protocols introduced in Refs. [48,49]

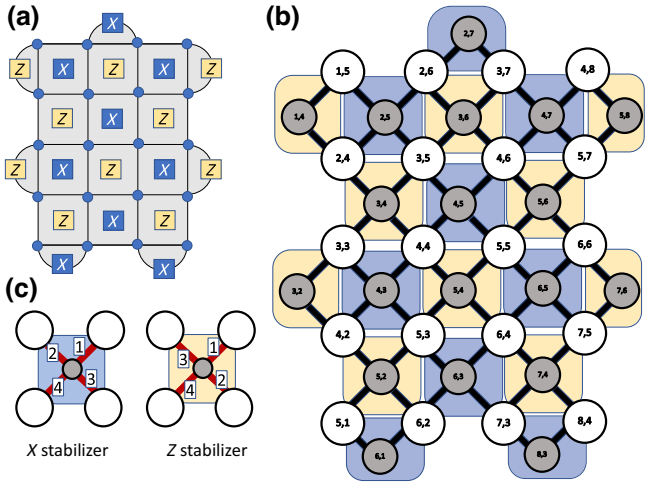

FIG. 1. (a) The standard schematic of a rotated surface code. In this case, we have a grid of 4×5 data qubits, where the data qubits live on the vertices of the code. The faces represent stabilizer measurements, in this case Z and X stabilizers. In the body of the code, the stabilizers are weight 4. Boundary conditions are dealt with by smaller weight-2 stabilizers as shown. (b) The realization of the code on the Sycamore device. The numbers in the circles identify the location of the qubit on the device grid. The ancilla qubits (in gray) reside in the center of each face. The data qubits are shown in white. The black lines represent the connections for two qubit gates that will be utilized to perform the circuits used to prepare the ancillas so that they can be measured to perform the stabilizer measurements. (c) In order to minimize the spread of errors, the ancillas need to be coupled to the data qubits in a very specific pattern. Here, we show the timing pattern of two-qubit gate activation for the ancillas used as Z and X stabilizers. In a complete stabilizer code implementation, the ancillas would be measured and reset after the completion of the stabilizer-preparation circuits. Here, we do not do this ancilla measurement (see text).

to extract comprehensive information about the noise in the device while the device is running the stabilizer-preparation circuits.

One of the key realizations we utilize is that two rounds of stabilizer preparation of the surface code perform an identity channel on the data qubits. This might not be immediately obvious given the complexity of the interleaved two-qubit gates but it is easily verifiable using a Clifford-circuit simulator. In Appendix C, we detail two rounds of stabilizer preparations.

In an actual surface-code circuit, the ancilla qubits would be measured and reset after each stabilizer-preparation round. However, where this is not yet possible or, indeed, where one wishes to examine the noise inherent in the circuits without introducing the additional noise that would be caused by measurement and reset, then the circuit extract shown in Appendix C is an example of the circuit required. One can also add measurements directly, although some care must be taken in this case (for more details, see Appendix C).

To eliminate coherent noise in the circuits, we introduce random Pauli gates into the circuit. These gates serve to randomize the Pauli frame, which on average turns the noise into a Pauli channel [54–56]. With Pauli frame randomization, the number of parameters describing the noise is, without further reductions, 4^d Pauli channel eigenvalues, where d is the number of data qubits.

If instead of an initial round of Paulis, we start every two-round block with a round of single-qubit Cliffords, then the analysis in Ref. [57] applies and we *locally average the noise* in the sense explored in Ref. [49] and summarized in Appendix E. This local averaging of the noise means that we only need to extract 2^d Pauli channel eigenvalues, which can be done in the course of this one experiment [49]. These 2^d parameters can be learned from extracting only the Pauli-Z-type eigenvalues of the Pauli noise channel and all of these eigenvalues can be determined simultaneously since they commute. The ability to locally average the eigenvalues is what makes this protocol highly efficient. The ability to interleave the local twirl with a Pauli twirl in between instances of the stabilizer-preparation circuits avoids the need to use multiqubit Cliffords to terminate the twirls.

With these adjustments in mind, the process follows the successful core idea of randomized benchmarking [58]: After initializing in a product state, we take each two-round block of stabilizer preparation, suitably averaged as above, and repeat it for m steps for varying lengths m. After an inversion step at the end, we measure in the computational basis. The data obtained in this way allow us to separate the state-preparation-and-measurement (SPAM) errors of the initial round from the errors of the "unit cell" of the two-round block of stabilizer extraction circuits. An example of this circuit for a single plaquette of the surface code is shown in Fig. 2. The full description of all of the steps required is set out in Appendix C and the analysis of the noise estimation is set out in Appendix D.

III. EXPERIMENTAL IMPLEMENTATION

We have tested these ideas on Google's Sycamore device, a 54-qubit superconducting device of the type described in Ref. [59]. In order to allow correct edge stabilizers, only 39 of the 54 qubits could be utilized, resulting in a 5×4 surface code (i.e., 20 data qubits), set out as shown in Fig. 1(c). For the Sycamore device, we used sequence lengths m (measured in double rounds of the surface code) of $m = [0, 1, 2, 3, 4, 6, 8]$. We ran 1770 different sequences, with 2000 shots per sequence. The total run time was around 8 h, in a dedicated 8-h time slot on the device.

Once we have the data, there are a number of things we can immediately do in postprocessing. The first is simply

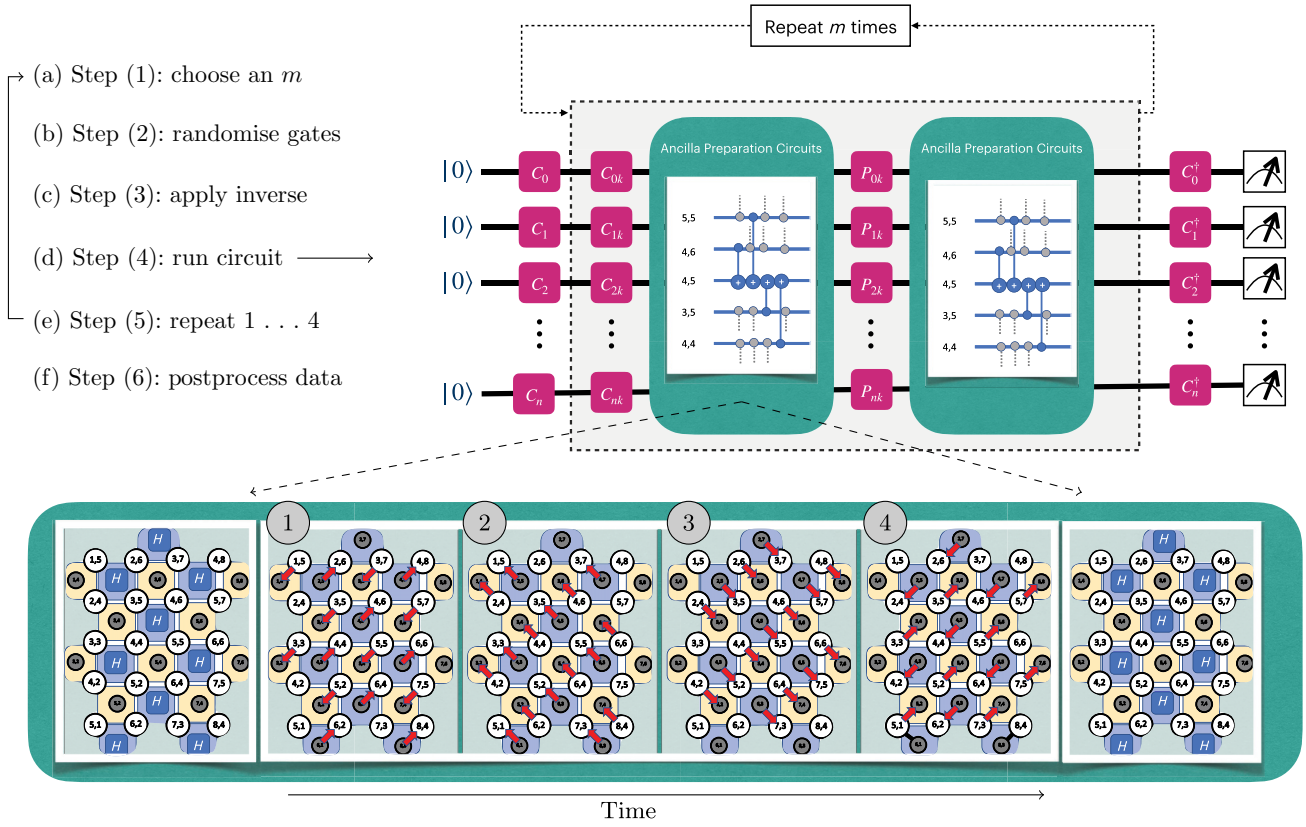

FIG. 2. A schematic of the data-acquisition step of the protocol. The boxes representing stabilizer-preparation circuits are one complete round of stabilizer preparation (without measurement or reset of the ancillas); one such round is detailed in the bottom schematic. The numbers (top left of the diagrams in the bottom schematic) represent the order of two qubit gates as shown in Fig. 1(c), the red arrows representing the two-qubit controlled-X (CX) gates. The large-H boxes (left and right subplots of the bottom schematic) represent Hadamard gates on the covered qubits. As with randomized benchmarking, the circuits are repeated for multiple sequences over multiple different lengths of circuit. The details of the steps are as follows. (a) We randomly choose a circuit length m from 0 (no stabilizer-preparation circuits) up to the maximum number of times the block in gray is to be repeated. The maximum value for m will depend on the errors in the device and should heuristically be chosen to be around the inverse of the average error rate of one block. (b) For an n-qubit system, randomly choose n starting Cliffords ($C_{0\ldots n}$) and $n \times m$ random Cliffords and Paulis [respectively, ($C_{0\ldots n,k}$), and ($P_{0\ldots n,k}$) for k in $\{1 \ldots m\}$]. These are used to locally average the noise. (c) The random single-qubit Cliffords and Paulis commute through the circuits in a way that is efficiently computable. Apply the inverse Clifford required to bring each qubit back into the computational basis. Note that we randomly determine whether a qubit is to be returned to the $|0\rangle$ or the $|1\rangle$ state and interpret the results so as to derive the chance of successfully returning to the desired state (i.e., record a 0 if the measurement accords with the desired state or a 1 if it indicates an error). (d) Run the circuit for a number of shots (2000 in our experiment) and record the resultant no-error or error bit patterns. (e) Steps (a)–(d) are repeated multiple times so that the results can be averaged. (f) Once the data are gathered, we can postprocess the results to obtain the required estimates. We describe the data-fitting process in more detail in Appendix C but, in brief, the data can be fitted to a model where we fit the data to a sequence of single exponential decays to estimate the relevant eigenvalues for the averaged circuit. The fitting to a decay curve makes the estimates independent of state-preparation-and-measurement (SPAM) errors.

to marginalize the data to each qubit and fit each such marginalization as per a standard one-qubit randomized benchmarking experiment. This gives us the single-qubit (and two-qubit) average error rates to high precision, with a maximum *relative* error of $\pm 0.1\%$. Second, if all one desired were the two-body correlations, then the data could be marginalized for each pair of qubits and the (normalized) correlations between two qubits (X and Y, $\rho_{X,Y}$), calculated as

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}}, \qquad (1)$$

where E is the expected value of the relevant qubit error rate. An example of these data directly extracted from the experiment is shown in Fig. 3. As can be seen in Fig. 3,

Qubit	Error
(1,5)	0.080
(2,6)	0.152
(3,7)	0.132
(4,8)	0.105
(2,4)	0.087
(3,5)	0.150
(4,6)	0.114
(5,7)	0.102
(3,3)	0.075
(4,4)	0.168
(5,5)	0.182
(6,6)	0.097
(4,2)	0.197
(5,3)	0.215
(6,4)	0.152
(7,5)	0.108
(5,1)	0.230
(6,2)	0.136
(7,3)	0.140
(8,4)	0.092

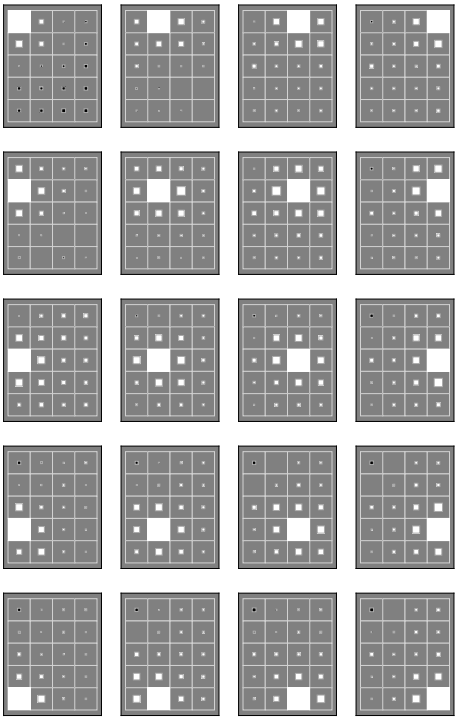

FIG. 3. Examples of some of the summary data that can be captured in a scalable manner directly from these types of experiments. By marginalizing the results, we can build up a table of the individual error rates, from one round of the surface code (left-hand side). These error rates represent the probability of any Pauli error occurring on the relevant qubit during one round of the stabilizer-preparation circuit. The right-hand side of the figure shows the two-body correlations between each data qubit (solid white in the appropriate subgraph) and each other data qubit in the code. The correlations between the errors on two qubits X and Y ($\rho_{X,Y}$) can be calculated in terms of moments as in Eq. (1). This Hinton diagram has a white square (black for negative correlations), where the area of the square is in proportion to the size of the correlation. Bootstrapping from the original measured sequences has been used to plot two-σ error bars, plotted as the width of the line bordering the square (they are barely visible in most cases). Of note is that the correlations between data qubits appear (mainly) to be stronger with local data qubits (i.e., those close in Manhattan distance according to the device connections). These local correlations are exactly the correlations we would expect in the surface code if we have noisy two-qubit gates (see Appendix D). There are, however, longer-range correlations that cannot be explained purely by noisy gates and must be caused by some form of crosstalk not accounted for in a circuit-level analysis.

there are correlated errors in the device that mostly cluster locally, although some significant longer-range correlations are present. While the experiment is only designed to identify and characterize errors rather than identify their cause, it is possible that the long-range two-body correlations noted in the figure and the multiqubit errors identified by the experiment are symptoms of the leakage errors in such Sycamore devices, as more fully explained in Ref. [60].

What is the impact of nonzero correlation on quantum error correction? If we know a particular qubit has an error (say, by a syndrome measurement), then the probability of the correlated qubit also having an error changes in a fashion that is computable, in expectation, by this correlation. A naive decoder assumes that the error rate on each qubit is independent of whether or not an error has occurred on a different qubit. Thus, these correlations can inform a decoder of the changed average error probabilities, improving average performance. The inclusion of such correlations is, in fact, necessary to obtain near-optimal performance [23,61].

However, we can do much more. Where the number of qubits is relatively low (no more than, say, 30 data qubits), then it is also possible to extract the entire probability distribution of errors in the device, which includes all larger-scale correlations. Here, the limiting factor is that the probability distribution will have 2^n elements and will rapidly become too large to store. The details in Appendix C show how to do this. Since we are interested in characterizing even larger-scale devices, we need to address this exponential growth in the full probability distribution. The experiment itself remains tractable but if we want to extract the full probability distribution, the postprocessing scales as $O(2^n)$. In principle, knowledge of this full distribution is required for optimal decoding but the prohibitive complexity makes obvious the need for approximations of this full distribution.

The rest of the paper explores ways to obtain a useful estimate of the noise without reconstructing the full probability distribution. For instance, graphical models of the distribution that represent the noise in a more compact form might give a sufficiently accurate description of the noise for decoding or other purposes, without such severe limits on complexity. In Sec. IV, we construct such models from our experiments and in Sec. V, we answer the question of whether such models allow faithful predictions of device performance.

IV. BUILDING AND TESTING NOISE MODELS

To go beyond brute-force descriptions of global probability distributions, we briefly review the notion of a graphical model, specialized for locally averaged Pauli noise. A locally averaged noise model on n qubits can be thought of as a probability distribution on n-bit strings, where the presence or absence of an error (either of X, Y, or Z) on the ith bit is denoted by a 0 or 1, respectively. Under the mild assumption that every error event has nonzero probability, this probability distribution can be described by a Hamiltonian $H(\mathbf{x})$ on bit strings $\mathbf{x}$, where the Hamiltonian can be chosen to be $H(\mathbf{x}) = \log p(\mathbf{x})$, the log of the probability of $\mathbf{x}$.

In this picture, an "energy shift" by a constant factor corresponds to changing the normalization of the probability distribution, which can often be neglected for sampling or modeling purposes. There is no need to invoke the physical concept of temperature since this is a formal mapping. We are interested in modeling a probability distribution as a graphical model the Hamiltonian of which has only a bounded number of interactions per qubit, each among a bounded number of other qubits, in order to keep learning tractable.

The simplest example of this mapping is where a single nonidentity Pauli error occurs with probability p and otherwise no error occurs. Then, $H(0) = \log(1-p)$ and $H(1) = \log(p)$. This Hamiltonian can be written as $H(x) = f + hx$, where $h = \log(p/(1-p))$ and where $f = \log(1-p)$ controls the (often optional) normalization. For n qubits and general independent noise, one would have the Hamiltonian $H(\mathbf{x}) = \sum_k h_k x_k$ (here and henceforth, we drop normalizations). The special case of identically distributed noise occurs when $h_k = h$ for all k. To generate interesting correlations, we need to add *coupling terms* to the Hamiltonian, and one is naturally led to models with Hamiltonians of the form

$$H(\mathbf{x}) = \sum_k h_k x_k + \sum_{j,k} J_{j,k} x_j x_k + \dots . \tag{2}$$

The simplest nontrivial case might contain only nearest-neighbor two-body correlations, for example.

For these models, learning the entire probability distribution is equivalent to learning the couplings of the associated Hamiltonian. Examples of such models are given in Fig. 4. In Fig. 4(a), we assume independent and identically distributed (IID) noise only (so that $H(\mathbf{x}) = h\sum_k x_k$). In Fig. 4(b), we keep independence but relax to nonidentically distributed (IND) noise, corresponding to $H(\mathbf{x}) = \sum_k h_k x_k$. Figure 4(c) shows the simplest model of nearest-neighbor correlation, an Ising-style model with $H(\mathbf{x}) = \sum_k h_k x_k + \sum_{\langle j,k\rangle} J_{j,k} x_j x_k$, where the second sum is over nearest neighbors. Finally, we consider a model that coarse grains the device into a one-dimensional (1D) array. This allows arbitrary correlations along one row of the device but limits vertical correlations to those between qubits in adjacent rows; this is shown in Fig. 4(d). The associated coarse-grained 1D (CG1D) model Hamiltonian has up to eight-body coupling terms.

The choice of which graphical models to examine is well motivated. The "Ising-style" model is relevant as the long-range spread of errors from one data qubit to another should be limited in the device, the gates coupling the data qubits to the ancillas acting as a type of "one-way" gate for Z and X errors [62]. Indeed, in Appendix H, we show how the data confirm that (for this device) the Ising model is a good choice compared with other potential decompositions of equivalent expressive power. Similarly motivated,

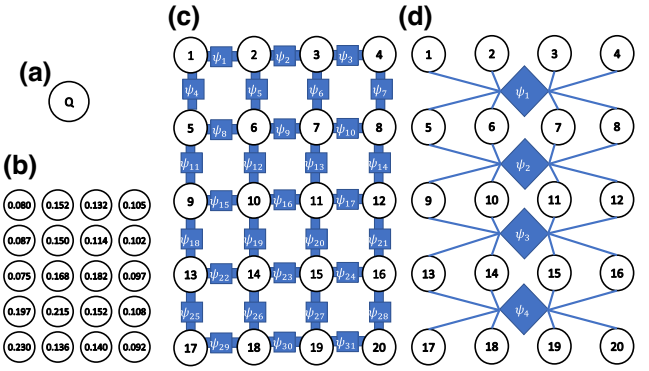

FIG. 4. Various graphical models used for modeling the observed error distribution. Sites connected by "factors" ψ_i can be correlated by arbitrary Hamiltonian couplings supported adjacent to the factors. (a) A typical model used for decoder testing. All the qubits are assumed to have identical independent depolarizing noise, so there is only one node. (b) Here, the qubits are still independent but each has their own error rate (here, the error rate shown in Fig. 3). This model requires n parameters, i.e., one for each data qubit—in this case 20. (c) Here, we model the noise with an Ising model where the only factors ψ_i are between neighboring qubits. (d) This model is referred to in the text as the CG1D model, where we have reduced the surface code to a one-dimensional (1D) graph.

the CG1D model is also able to capture longer correlations, albeit at the cost of exponential scaling in the width of one dimension. While problematic for larger systems, it might still be a valid methodology in systems with highly biased noise, where the proposed grid would be rectangular rather than square [63,64]. Here, as our device has a slight asymmetry (a height of 5 and a width of 4), we have chosen to use the width dimension—although in this particular case nothing turns on this choice save for the computational complexity of modeling and sampling the network.

While these models attempt to match the correlation structure in the device, some model error is inevitable. For instance, the Ising model assumes local interactions (which we imagine arising from gate errors) but there are many other noise mechanisms that can cause errors to spread from one qubit to another. Examples include energy-state leakage [60], leakage of control signals, and qubit frequency crowding. Model error in this case will indicate such processes at work.

One method of measuring model error is to use the relative entropy or, more flexibly, the smoothed symmetrized version of it, the Jensen-Shannon divergence (JSD) [65]. One can measure the JSD as between the global distribution and the distribution encapsulated in the model to quantify the model error. Here, we have all the data required to make this calculation. However, to do so in general requires the full distribution and therefore this is not a scalable solution. Another potential method might be to form the covariance matrices embodied in each of

TABLE I. The metrics for the observed error distribution (D) compared with the distributions embodied in the various models (M). The two error metrics are the norm between the respective correlation matrices and the Jensen-Shannon distance (JSD) (see the main text for definitions). The models are identical independently distributed (IID) errors [Fig. 4(a)], independent nonidentically distributed (IND) errors [Fig. 4(b)], Ising, with nearest-neighbor two-qubit correlations only [Fig. 4(c)], and the coarse-grained 1D (CG1D) distribution shown in Fig. 4(d).

Model	JSD ($D\|M$)	$\|\Sigma_D - \Sigma_M\|$
IID	0.229	0.124
IND	0.192	0.090
Ising	0.167	0.056
CG1D	0.148	0.019

the distributions—this can be done in a scalable manner. (Obviously, the models shown in Figs. 4(a) and 4(b) have no interqubit correlations.) These covariance matrices only capture the two-body correlations in the distributions and technically only form a lower bound on the total variational distance (TVD) of the distributions (Appendix G) but are usable as a guide in most nonpathological cases. We set out these calculations in Table I. Unsurprisingly, as can be seen from Table I, the richer the model, the more complexity we can capture in the correlations between the qubits.

The question then arises not as to whether these more sophisticated models can capture characteristics of the noise that cannot be captured by the simpler models but whether those additional characteristics are important in determining and/or predicting error-correction properties in the device. We now address this question, focusing on the logical error rate of the device.

V. LOGICAL-ERROR-RATE ANALYSIS

In order to generate counterfactual error distributions that are less noisy but still contain similar correlation structures, we consider the following one-parameter family of error channels associated with a probability distribution p on bit strings. The error probabilities p can be related to the eigenvalues λ of the superoperator representation of the channel using W, the Walsh-Hadamard transformation [48,49,67]. In fact, we have $\lambda = Wp$ and this map is invertible to obtain $p = W^{-1}\lambda$. If this superoperator with eigenvalues λ was generated by a continuous-time process, then we could generate the family of channels with eigenvalues

$$\lambda(t) = \exp(t \log \lambda(1)), \quad \lambda(1) = Wp. \tag{3}$$

When $t = 0$, this is the identity channel and $t = 1$ gives back the original noise-channel eigenvalues. We can study the error rates of this distribution by considering

$$p(t) = W^{-1}\lambda(t) \tag{4}$$

for various values of $t \geq 0$, with smaller values corresponding to less noise but with a qualitatively similar correlation structure as the true channel $p(1)$. Note that larger integer values $t = 2, 3, \ldots$ correspond to multiple applications of noise.

There is an important subtlety: not every input p is "divisible" into a continuous-time process, even in the simple case of Pauli noise channels. We therefore add an extra step of projecting $p(t)$ defined by Eq. (4) to the nearest point on the probability simplex. For more discussion and analysis of the probability distributions obtained by this method, see Appendix I.

To summarize: our experiment gives us the observed distribution of the actual noise in the device and our graphical models have given us estimates of this noise in a convenient and scalable format. By making the additional assumption that the noise is generated by a continuous-time process, we use Eqs. (3) and (4) to generate counterfactual theoretical noise distributions that are *estimates* of the noise that would exist in the device if the noise channel were "less" noisy but retained the correlation features of the observed noise channel. We can also construct models of this counterfactual noise.

The following calculations of the logical error rate are based on two additional important assumptions. First, as mentioned previously, this experiment omits the measurement and reset of ancillas required for true quantum error correction. These will, undoubtedly, introduce complexity, noise, measurement errors, and timing issues in any final circuit. The numbers obtained and discussed below must be regarded with this caveat in mind. Second, by using a generic decoder, we are not attempting to utilize our knowledge of the noise to *improve* the decoding process. The decoder utilized is a generic decoder provided by QECSIM [66]. Finally, we note that we only have knowledge of the locally averaged errors (e.g., for a particular qubit, we only know the average of the Pauli-X, -Y and -Z errors)—not their individual rates. For the purposes of the discussion below, we have assumed that where an error occurs, it is equally likely to be any one of these three Pauli errors. If one wanted to accurately predict the logical error rate for a particular device, such knowledge could be important.

While there is much work yet to be done to work out if actual knowledge of the noise can improve decoding success rates [31], this is not what we do here—rather, the analysis below might be seen as setting out the base success rate, which such decoders might seek to improve on. Writing such decoders is the subject of ongoing work.

With this in mind, we can now calculate the logical error rate of both the observed and constructed noise channels and we can determine if the more sophisticated models of that noise are better able to predict the likely logical error rate than their primitive counterparts. If so, then it is likely

Model	Predictions	
	Threshold	Logical error if physical error averages 0.031
Ising	0.0069	0.0191
CG1D	0.0098	0.0156
Global	0.0105	0.0127
IND	0.0114	0.0068
IID	0.0114	0.0066

FIG. 5. A plot of the logical error rate of various physical error distributions, where we have used a generic QECSIM [66] decoder (a tensor network decoder with bond dimension $\chi = 8$) to decode the error syndromes. The process involved is as follows. (1) Measuring the empirical global distribution using the procedures discussed in this paper. (2) Constructing counterfactual global error distributions with smaller error rates (*reconstructed distributions*). The methodology for this step is discussed in the text and in Appendix I. (3) For each of these distributions (observed and reconstructed), each model of interest (see Fig. 4) is constructed. (4) The relevant distributions and each of the models are then sampled to provide a sample of the errors given by each of them. For each sample, we use the QECSIM decoder to determine whether a successful decoding of the error syndrome would be made. (To enable the sampled locally averaged errors to simulate a full Pauli distribution, we assume an Pauli-X, -Y, or -Z error, with equal probability for every affected site.) (5) Sufficient samples are taken of each distribution and each model of that distribution to allow the estimation of a logical error rate for that distribution or model of the distribution. (6) Steps (4) and (5) are repeated 10 times to generate the reported bootstrap error bars.

that they are capturing essential elements of the noise that the simpler models cannot.

Figure 5 shows the results of these logical-error-rate estimates for the different noise models. In each case, 10 000 error samples were taken to estimate the logical error rate, repeated 10 times to provide the error bars. Of interest in the plot is the behavior when the average per-qubit physical error rate is approximately 0.136 [point (a) in the plot]. This is the measured noise in the device. At that point, the logical error rate is larger than this (0.176) but notably there is very little difference in the logical error rate between the various models (from the most primitive IID model to the CG1D distribution) and the full observed distribution. At this point in the spectrum of noise channels, the simple IID noise model is as good as any [point (a) in the figure].

However, as we use higher-fidelity reconstructed distributions, a clear difference emerges. By the time we hit point (b) in the figure, the simpler models result in an estimated logical error rate that is significantly lower than the logical error rate predicted by the full reconstructed distribution. The correlated errors have an impact in this regime, curtailing the ability of the generic decoder to correct for errors. (Again, we note that a decoder tailored for these noise characteristics might perform much better.) This accords with beliefs that correlated noise in a machine will have an impact on logical error rates. Interestingly, even in this regime, the CG1D model is within error bars of predicting the same logical error rate as the full reconstructed distribution.

The Ising model is pessimistic in its predictions (predicting a higher logical error rate). Our belief is that this is due to the Ising model not being able to correctly capture longer-range correlations that happen to exist in this device. The parametrization of the Ising model therefore incorporates these longer-range correlations into the short-range correlations it can model, resulting in a model that has stronger short-range correlations than the full distribution.

In this small (low-distance) surface-code implementation, it is these stronger short-range correlations that have the largest impact on the logical error rate. Despite this

limitation, as we show in Appendix H, if we constrain the model to use only a limited number of two-qubit factors, the Ising model is still the optimal model of such distributions, minimizing model error (and possibly avoiding overfitting). On these data, one could be confident that if an Ising model of the distribution, populated from data from the device, was below threshold, the actual device would also be likely to be below threshold.

VI. CONCLUSIONS

We have presented and implemented a method of characterizing an important part of the surface code, namely, the stabilizer-preparation portion of an error-correction circuit. We have shown how one can use random benchmarking-style experiments to measure the locally averaged noise and the Pauli noise on the data qubits used in the code. We have shown how to create graphical models of the noise that continue to be tractable as surface-code sizes increase. These models allow us to explore important questions related to the ability to run error correction on the device. Finally, we have presented empirical evidence that shows how these more sophisticated models appear to be increasingly necessary if one wishes to accurately predict the errors in the device as error rates get lower and correlations shorter range. The importance of taking such errors into account is highlighted by the data that we extrapolate from the device, providing support to the belief that decoders that take into account the actual noise of the device could potentially lead to higher threshold implementations of error-correcting codes.

All code and data are available upon reasonable request.

ACKNOWLEDGMENTS

The experiment was performed in collaboration with the Google Quantum AI hardware team under the direction of Y. Chen, J. Kelly, and A. Megrant. We acknowledge the work of the team in fabricating and packaging the processor, building and outfitting the cryogenic and control systems, executing baseline calibrations, optimizing processor performance, and providing the tools to execute the experiment. We thank D. Debroy, B. Foxen, M. Harrigan, and M. Newman for reading the manuscript thoroughly and providing helpful feedback. R.H. would like to thank Robin Blume-Kohout for several insightful conversations relating to the paper. R.H. is funded by the Sydney Quantum Academy and this work was supported by Army Research Office (ARO) Grant No. W911NF2110001. S.F.'s contributions to this project were completed while he was affiliated with the University of Sydney.

APPENDIX A: NOISE CHARACTERIZATION

Many characterization techniques have been developed to understand noise in quantum devices. Traditionally, full-characterization techniques such as process tomography have been used when devices are in their infancy. For small devices, the qubits can be fully characterized with a view to gaining insight into the underlying causes of noise. Process tomography [33] and its variants [34–36] remain important tools in the characterization arsenal. However, such methods are limited in that they can only be applied to a small number of qubits. Where we have more than a few qubits, scalablity issues render the techniques impractical.

Strong noise averaging or partial characterization is a way to overcome the scalability issues of full characterization. A popular partial-characterization technique is randomized benchmarking and its variants [37–46]. These techniques define a natural measure of a "block" of noise over which the noise in the device is averaged. By defining a block of noise that is well behaved through multiple applications, simple decay curves can used to estimate parameters in a way that eliminates SPAM errors and provides small error bars [68]. Importantly, these protocols average the noise in the device and this multiple-run average turns the noise into the equivalent of a Pauli channel, removing coherent effects of the noise. Depending on the protocol, the Pauli noise itself may be additionally averaged, often reducing the noise to a single value, the fidelity of the overall channel. The conversion to a Pauli channel is justified as the noise in the device can itself be converted into a Pauli channel during the operation of a device by using techniques such as Pauli frame randomization [54–56]. The subsequent strong averaging is a characterization convenience.

Rather than seeking to measure the average noise of the entire Pauli channel, further advances have shown how to harness the power of simultaneous single-qubit measurements to extract much more information about the noise [48,49,69–73]. Recently, it has been shown that one can estimate Pauli noise channels from error-correction syndrome measurements [74], although in the case of surface-code style error correction, this is limited to a maximum of two-body correlations, irrespective of the size of the code. Here, we will use techniques that allow the extraction of the global error distribution [48,49] on devices running error-correcting circuits and show how to tame this exponentially large amount of data by constructing appropriate models of the noise.

APPENDIX B: DEVICE CHARACTERISTICS

The device used in these experiments is the exactly the same 54-qubit Sycamore processor as in Ref. [59]. We operate 39 qubits selected from a 53-qubit grid, with idle and interaction frequencies chosen to be optimized for on-resonance $\sqrt{(i\mathrm{SWAP})}$ and Sycamore gates; while running circuits on these 39 qubits, the unused qubits in the 53-qubit grid idle at their normal operating frequencies and are not biased to low frequencies. Automated calibrations

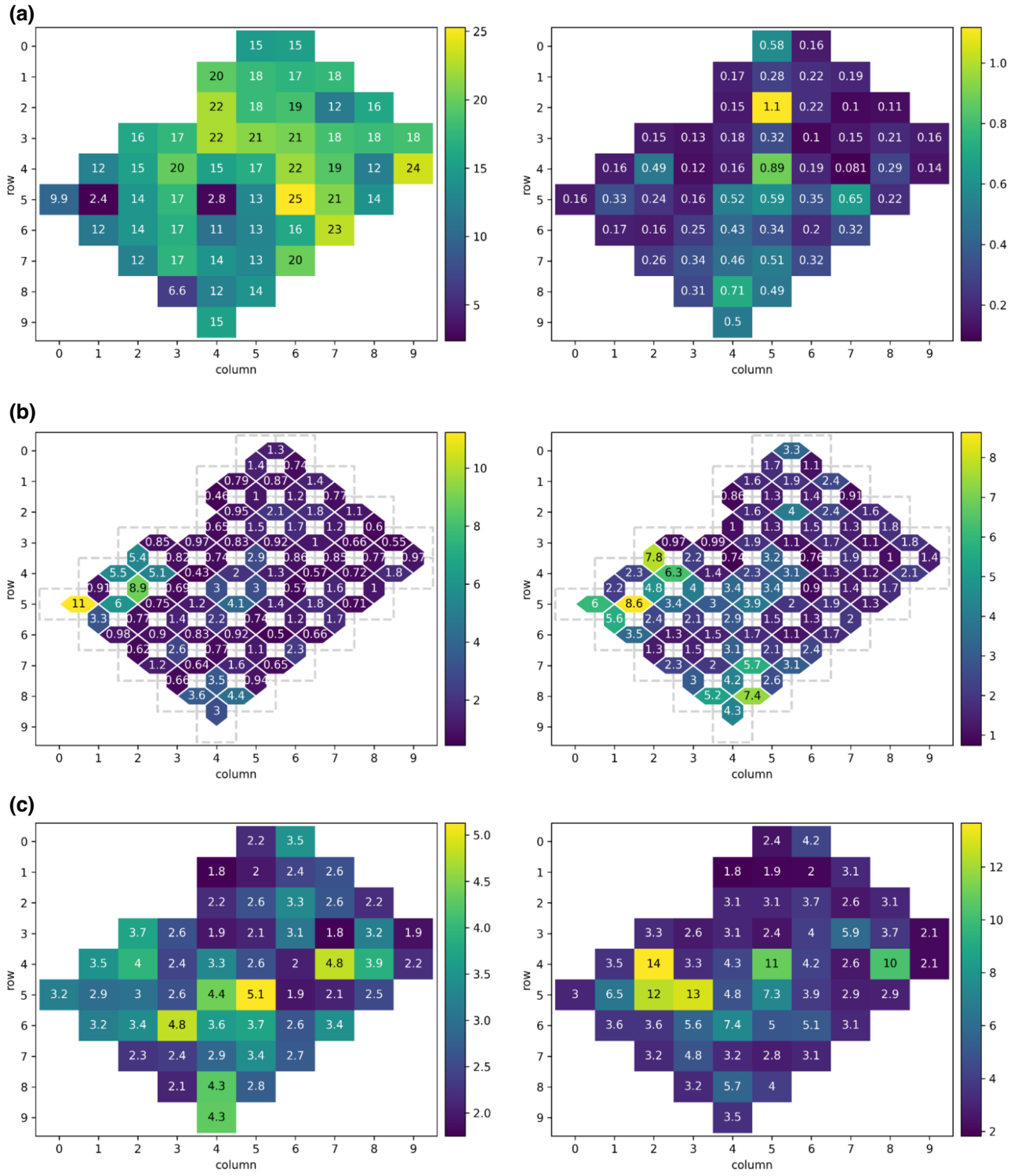

FIG. 6. (a) Typical (left) qubit T_1 lifetimes (in μs) and their frequency and (right) single-qubit ("1Q") cross-entropy benchmarking (XEB) Pauli errors (×100). (b) Isolated (left) and parallel (right) two-qubit ("2Q") XEB Pauli errors per cycle (×100). (c) Isolated (left) and simultaneous (right) measurement errors (×100). In all cases, for further details, see Appendix B.

were performed and control parameters updated just three times a week with minimal manual intervention. Instabilities in coherence times and electronics drift were not compensated for in the intervening time. Most data presented in this paper were acquired approximately 60 h after the previous calibration and control-parameter update.

Immediately following each automated calibration cycle, gate and readout fidelities were characterized to provide representative device-performance characteristics as summarized in Fig. 6. Figure 6(a) shows typical qubit T_1 lifetimes at their idle frequencies and Fig. 6(b) plots single-qubit Pauli errors for $\pi/2$ gates characterized with isolated cross-entropy benchmarking (XEB), where isolated operation is defined as applying gates to a single qubit with the remaining 52 idling at their idle frequency. On this processor, we have observed only a modest increase in single-qubit gate errors during simultaneous operation (see Ref. [59, Fig. 2])but significant differences are present for isolated and simultaneous readout and two-qubit gate operations. Since the precise mechanisms of crosstalk are not well understood, we assume that performance is dependent on exactly which set of qubits are operated simultaneously. To provide a representative estimate of these crosstalk effects, in Figs. 6(b) and 6(c) we provide isolated and simultaneous measurements of Sycamore gate fidelity and readout error. Note that Sycamore gate fidelity is characterized in four layers to characterize gates using each of the (up to) four couplers connected to each qubit. Finally, we note again that the procedure we describe is robust to SPAM errors, although high SPAM errors will reduce the signal and therefore increase the number of measurements needed to achieve the same accuracy.

While Appendix C describes the circuits run in terms of Clifford gates, natively the Sycamore device runs Sycamore gates for its two-qubit gates. The controlled-X (CX) gates are translated into two rounds of Sycamore gates

$$\mathrm{SYC} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 \\ 0 & -i & 0 & 0 \\ 0 & 0 & 0 & e^{-i\pi/6} \end{bmatrix},$$

with appropriate single-qubit rotations. The single-qubit rotations (being the random single-qubit Cliffords, the Sycamore gate corrections, the random Paulis, and the required Hadamard gates) can also be combined into single-phased XZ gates, defined as

$$PhXZ(x,z,a) = \begin{bmatrix} e^{i\pi x/2}\cos(\pi x/2) & -ie^{i\pi(x/2-a)}\sin(\pi x/2) \\ -ie^{i\pi(x/2+z+a)}\sin(\pi x/2) & e^{i\pi(x/2+z)}\cos(\pi x/2) \end{bmatrix}.$$

This is done automatically by using the appropriate CIRQ functions to optimize the circuits for the Sycamore device.

An extract of the circuit for a particular selection of random Cliffords and Paulis is shown in Fig. 7(b).

APPENDIX C: DESCRIPTION OF THE EXPERIMENTS

As with most SPAM-robust protocols that measure incoherent noise rates, the experiment design utilizes many of the features of randomized benchmarking. We use single-qubit Clifford gates to locally average the noise, with additional Pauli frame randomization where we cannot easily use Clifford gates to remove coherent noise between iterations of the surface-code circuits.

An overview of the design and the related circuit extract is shown as Fig. 7. The full procedure is as follows:

(1) Choose a non-negative integer m.
(2) For each qubit, randomly decide whether to leave the state invariant or to map it to an orthogonal state in order to eliminate a nuisance-model parameter (see, e.g., Refs. [68,75,76]).
(3) Create m applications of the following circuit:
    (a) random single-qubit Clifford gate on each qubit
    (b) apply one round of the stabilizer-preparation circuits
    (c) apply a random Pauli gate on each qubit
    (d) apply one round of the stabilizer-preparation circuits

    Note that if m is 0, then still apply the single-qubit Clifford gate and the inverting gate in step (4).
(4) Choose the random single-qubit Clifford gate needed to return each particular qubit to the state chosen in step (2).
(5) Run the circuit created in steps (2)–(4) a number of times (in the experiment, we chose 2000), measuring all the qubits. Record the bit patterns from the measurements.
(6) Repeat steps (1)–(5) for various values of m. In the current experiment, we chose $m \in [0, 1, 2, 3, 4, 6, 8]$. Reference [68] provides guidance but, in general, you want the largest m to be large enough that most qubits have a marginalized survival rate that is approximately 60%.
(7) For each distinct m chosen, repeat steps (1)–(6) sufficient times to obtain reasonable statistics (in the reported experiment, we ran each m over 1700 times but far fewer runs are actually required for reasonable error bars).

If the number of data qubits (n) is such that the 2^n is a tractable number:

(1) For each m, marginalize the bit patterns to the data qubits and use these reduced bit patterns to create a 2^n-outcome empirical probability distribution. It is not a concern that not all possible bit patterns will have been observed. These values are 0 in the empirical probability distribution.
(2) Walsh-Hadamard transform each of these m probability distributions, forming m eigenvalue vectors (each eigenvalue vector having 2^n entries).

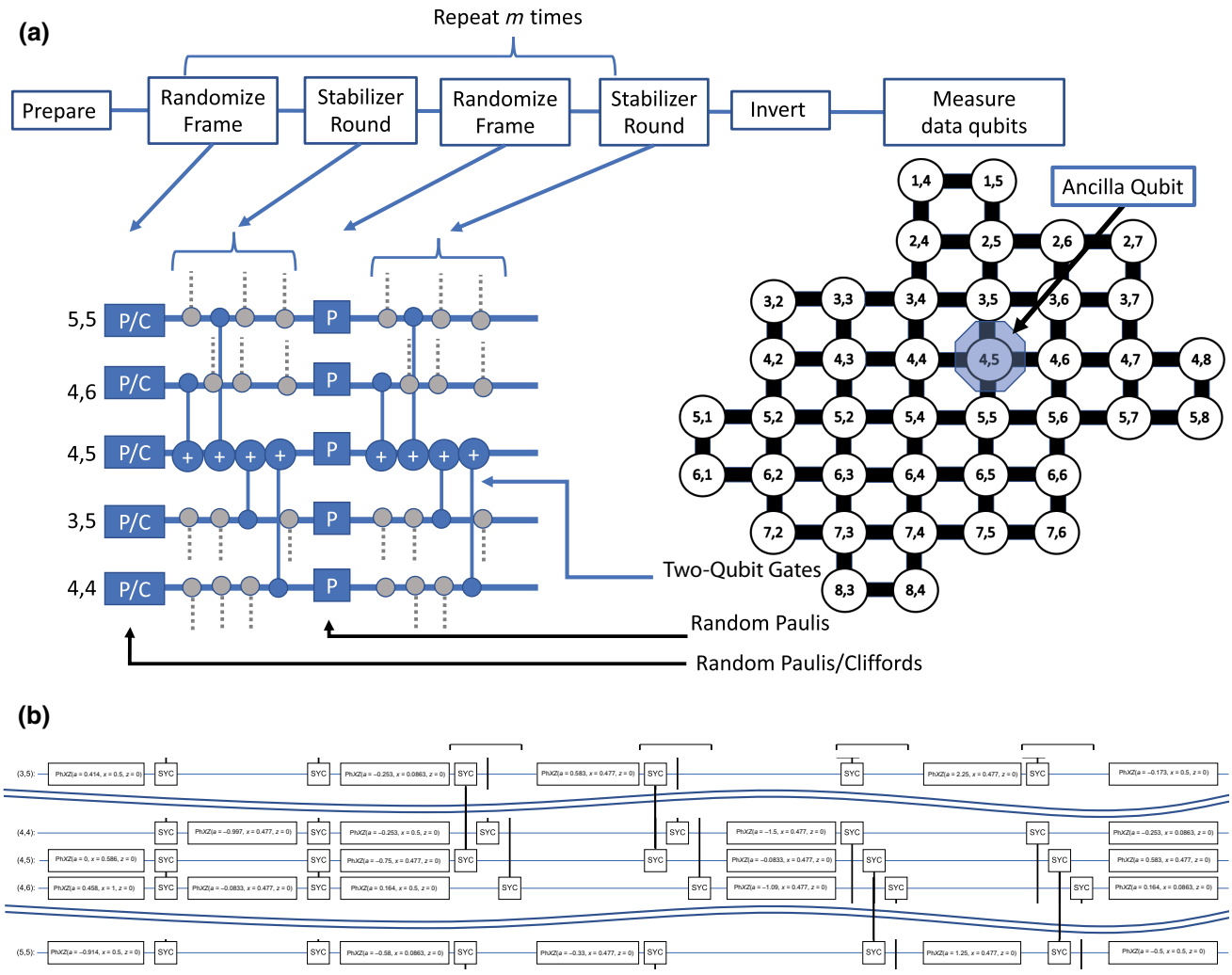

FIG. 7. (a) An example of the circuits that we measure. These are the same circuits as used to prepare the ancilla qubits, so as to allow syndrome extraction. We show only the gates for one ancilla (qubit 4,5). (We have included "shadows" of other gates just to remind the reader that many other two-qubit gates are being executed simultaneously with the gates shown; these "shadow" gates are in gray.) The inclusion of random Pauli gates ensures Pauli frame randomization, which when averaged over many runs means that the statistics gathered are as if the noise channel we are measuring were a Pauli noise channel. If we replace the first round of Pauli gates with random single-qubit Clifford gates, then we can locally average the noise in a sense made more precise in the text. The circuit extract shown here is a round of gates. Each round returns the qubits to the computational basis, subject to a random single-qubit Pauli (or Clifford) on the qubits. This is easily tracked and inverted just prior to readout. By preparing and measuring sequences for circuits with varying lengths (i.e., varying m in the above graphic), we are able to transform observed probabilities into eigenvalues and fit to a decay curve (see Refs. [48,49]). Reference [48] proves the convergence properties of such circuits to estimate the probability distribution of the average noise in the system. In this case, the noise measured is the average noise while running exactly the type of circuit in which we are interested. (b) The native two-qubit gates on the device are the Sycamore gates. Two Sycamore gates (and some single-qubit gates) are required to implement a CX. Here, we show most of one round of the above circuit extract (with some randomly chosen single-qubit Clifford gates), detailing the gates actually executed on the device. Where a Sycamore gate connects to a qubit not shown in this extract, only a single leg is shown.

(3) For each of the m eigenvalue vectors, for i in $2 \ldots 2^n$, extract the ith eigenvalue ($\bar{\lambda}_i$) and fit to the following equation:

$$\bar{\lambda}_i(m) = A f_i^{\,m}. \tag{C1}$$

Record each f_i from the fitting procedure.

(4) Form a new eigenvalue vector $[1, f_2 \ldots f_{2^n}]$. If desired, Walsh-Hadamard transform this to a probability distribution and project onto the nearest point in the probability simplex.

If 2^n is not a tractable number, marginalize the bit patterns to tractable chunks and use the resultant marginal

probability distributions to determine the parameters of a chosen model. Connection of the (empirical) marginal probabilities to the model parameters of an undirected graphical model to the model parameters is discussed at length in Ref. [48, Sec. VII] and we describe how to do this parameter estimation for our models in more detail in Appendix F.

Measurements of the ancilla qubits can be included by expanding the circuits where it is desired to include ancilla measurement and reset. Some care is required, however, as such measurement and reset must only be done after two rounds of stabilizer preparation (unlike actual execution of the surface code, where measurement and reset occurs after each round of stabilizer preparation). Performing the experiment on a circuit without measurement and reset and then with measurement and reset will allow the noise associated with the measurement and reset to be extracted in a similar way to interleaved benchmarking.

APPENDIX D: ANALYSIS OF THE DATA

The main idea behind applying the analysis to stabilizer-preparation circuits is that we can leverage the self-inverting nature of a round of these circuits to implement the protocol described in Ref. [57]. As stated in that paper, the protocol is applicable to any "gate" A, where A^2 and $APA^\dagger$ are elements of the Clifford group (here, P is a Pauli). In this case, a round of stabilizer-preparation circuits, which can be thought of as a large multiqubit gate (being composed of Clifford gates), satisfies those conditions. In more detail, let $\mathcal{S}$ represent a round of stabilizer-preparation circuits and let $\Lambda_{\mathcal{S}}$ represent the noise on such a round. We use a tilde ("$\sim$"), as in $\widetilde{\mathcal{S}}$, to represent a noisy "gate"; here, a noisy implementation of a round of stabilizer-preparation circuits. Given this, we have

$$\widetilde{\mathcal{S}} = \mathcal{S}\Lambda_{\mathcal{S}} \tag{D1}$$

$$\mathcal{S}\mathcal{S} = I \implies \mathcal{S} = \mathcal{S}^\dagger \tag{D2}$$

where, in the first line, we have arbitrarily written the noise on the on the right-hand side of the gate. (Nothing depends on this.)

We use C_j to represent a series of n single-qubit gates, the suffix representing the jth draw of a set of single-qubit Cliffords and n being the number of qubits used to implement a round of the surface code (both data and ancilla qubits). $\mathcal{C}$ represents the set of all possible instantiations of C and $|\mathcal{C}|$ is the size of this set. Similarly, P_j represents n single-qubit Paulis and $\mathcal{P}$ and $|\mathcal{P}|$ have analogous meanings.

In this notation, a round of the protocol looks like

$$\langle 0|^{\otimes n} C_{\text{inv}} \dots C_{j+1}\tilde{\mathcal{S}}P_{j+1}\tilde{\mathcal{S}}C_j\tilde{\mathcal{S}}P_j\tilde{\mathcal{S}}\dots|0\rangle^{\otimes n}. \tag{D3}$$

There are three further things to note. (1) We have ignored the noise on the rounds of Paulis and single-qubit Cliffords, they will be trivial compared to a full round of the surface-code and stabilizer-preparation circuits and, where possible, the transpiler will compile them into the physically realized gates. (2) Paulis commute through a round of the stabilizer-preparation circuits into other Paulis (since it is comprised of Clifford gates), i.e.,

$$P\mathcal{S} = \mathcal{S}P'. \tag{D4}$$

(3) The sequence $P\mathcal{S}$, being a round of perfect Paulis followed by a perfect round of stabilizer-preparation circuits, is a unitary-1 design. This means that conjugating a noise channel with $P\mathcal{S}$, averaged over all Paulis, will remove the coherence in the noise, reducing it to a Pauli channel, i.e.,

$$\frac{1}{|\mathcal{P}|}\sum_{P\in\mathcal{P}} P\,\tilde{\mathcal{S}}\,P'\,\mathcal{S} = = \frac{1}{|\mathcal{P}|}\sum_{P\in\mathcal{P}} P\,\mathcal{S}\,\Lambda_{\mathcal{S}}\,\mathcal{S}\,P \tag{D5}$$

$$= \frac{1}{|\mathcal{P}|}\sum_{P\in\mathcal{P}} (P\,\mathcal{S})\,\Lambda_{\mathcal{S}}(P\,\mathcal{S})^\dagger \tag{D6}$$

$$= \Lambda_{P\mathcal{S}}, \tag{D7}$$

where in Eq. (D5), P' has been chosen so that it commutes through $\mathcal{S}$ to become P and in Eq. (D6), we have used the fact that for both the Paulis and for a round of the stabilizer-preparation circuits they are their own inverse [see Eq. (D2)]. Here, $\Lambda_{P\mathcal{S}}$ represents the Pauli-twirled noise channel on a single round of the surface code, being a Pauli channel. Note that although the Pauli twirls themselves do not permute the eigenvalues, the stabilizer-preparation circuits (S) will. Because the circuits contain two-qubit gates, these gates, if noisy, will serve to spread errors between qubits—which will be seen as correlations in the protocol. We discuss this more at the end of this appendix.

With this is hand, we can now analyze the proposed protocol. The repeating portion of the protocol is as follows:

$$C_{j+1}\tilde{\mathcal{S}}\,P_{j+1}\,\tilde{\mathcal{S}}\,C_j = C_{j+1}\,\mathcal{S}\,\Lambda_{\mathcal{S}}P_{j+1}\mathcal{S}\,\Lambda_{\mathcal{S}}\,C_j. \tag{D8}$$

Now, we can let P' be defined as $P_{j+1}\mathcal{S} = \mathcal{S}P'$ and define C'_{j+1} as $C_{j+1} = C'_{j+1}P'$, noting that if C_{j+1} was chosen randomly, C'_{j+1} is also a random Clifford. Then,

$$C_{j+1}\,\mathcal{S}\,\Lambda_{\mathcal{S}}P_{j+1}\mathcal{S}\,\Lambda_{\mathcal{S}}\,C_j \tag{D9}$$

$$= C'_{j+1}\,P'\,\mathcal{S}\,\Lambda_{\mathcal{S}}P_{j+1}\mathcal{S}\,\Lambda_{\mathcal{S}}\,C_j \tag{D10}$$

$$= C'_{j+1}\,(P'\mathcal{S})\,\Lambda_{\mathcal{S}}\,(\mathcal{S}P')\,\Lambda_{\mathcal{S}}\,C_j. \tag{D11}$$

Finally, we can write C'_{j+1} as $C''_{j+1}C_j^\dagger$:

$$C''_{j+1}(C_j)^\dagger\big[\,(P'\mathcal{S})\,\Lambda_{\mathcal{S}}(\mathcal{S}P')\Lambda_{\mathcal{S}}\,\big]C_j, \tag{D12}$$

which gives us a Pauli twirl of the noise, embedded in a Clifford twirl of the noise (when averaged over sufficient sequences). The exact nature of this twirl is analyzed in detail in Ref. [57]. A question might be asked as to why we should not just use rounds of single-qubit Cliffords to locally average the noise, rather than the alternating Clifford and Pauli rounds analyzed above. The reason is that while single-qubit Cliffords can indeed be commuted through an odd number of rounds of the surface code, the resulting Clifford is no longer in the group of simultaneous single-qubit Cliffords but, rather, is a number of multiqubit Clifford gates. Chasing this through the circuits results in a multiqubit inversion gate at the end, which will substantially reduce the signal being measured.

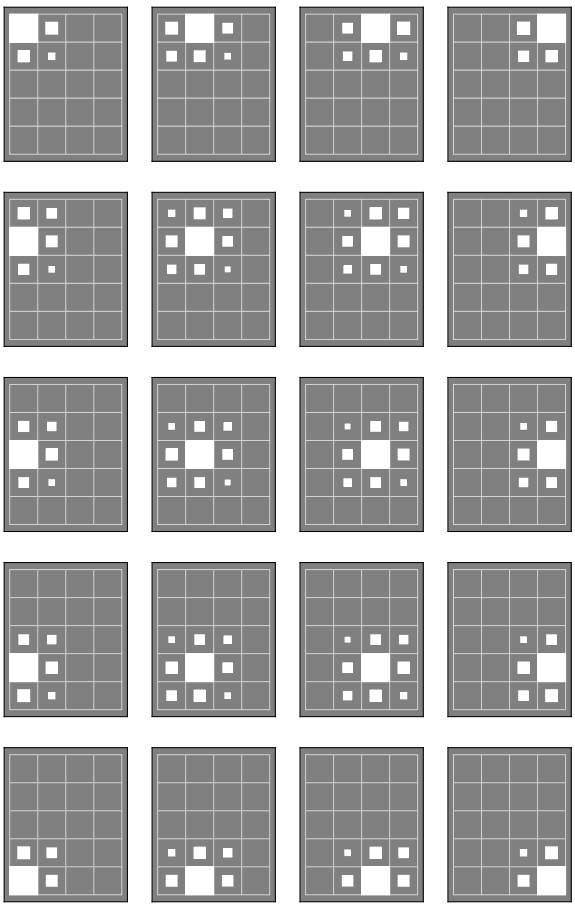

FIG. 8. Expected correlations on a 4×5 surface code. The simulation was carried out using STIM [77] with depolarizing noise on the two-qubit gates and the single-qubit gates and X-flip noise on the measurements. Only the data qubits are shown. As before, The full white square represents the qubit being examined in the each subplot. The data qubits are laid out in the same "grid"-like pattern in which they appear in the surface code, i.e., neighboring data qubits are next to each other. For any particular qubit, the "X" ancilla could correlate errors between it and the three other data qubits it touches. In addition, the "Z" ancilla could spread correlations between that qubit and up to three other data qubits. Consequently, we would not expect to see any data qubit have correlated errors with a nonadjacent qubit. That is the case.

Finally, we look at the expected correlations if the stabilizer-preparation circuits are carried out using noisy two-qubit gates that otherwise behave as per the circuit model. These circuits are designed to be fault tolerant—i.e., to limit the spread of errors. Errors can spread from data qubits to ancillas (X or Z depending on the plaquette) and from ancillas to the surrounding data qubits (Z or X—again depending on the plaquette). This should result in no data qubit having correlated errors other than with respect to the (up to) eight other data qubits with which it shares ancillas. If we simulate the protocol using STIM [77], we see this expected pattern (Fig. 8).

APPENDIX E: LOCALLY AVERAGED NOISE CHANNEL

The meaning of a locally averaged noise channel is fully explored in Ref. [49], which builds on the proofs in Ref. [48]. However, for the sake of completeness, we summarize the results here. It is well known that averaging observed distributions over sequences of gates drawn from particular groups eliminates coherence and will average the noise in a way that is dependent on the groups in question. This is colloquially known as "twirling" the noise. Reference [43] explores this in detail. If the group in question is the Pauli group, then the noise being twirled, when averaged over repeated randomized sequences, is the equivalent to a Pauli noise channel, i.e., the noise with all of its coherence removed. When the group is the full Clifford group, then the noise becomes a quantum depolarizing channel with the same fidelity as the original channel [78]; this is one of the essential ingredients of randomized benchmarking [39]. Twirling the noise with single-qubit Clifford gates averages the noise, such that the number of distinct eigenvalues of an n-qubit noise channel is reduced from the 4^n distinct eigenvalues of a Pauli channel to 2^n distinct values. It does this by averaging the Pauli noise locally. For example, in a two-qubit channel, the Paulis IX, IY, and IZ are averaged, the Paulis XI, YI, and ZI are averaged and the remaining nine two-qubit Paulis are averaged. In this case, the four distinct eigenvalues (2^2) can be recovered by measuring the eigenvalues of the Paulis II, IZ, ZI, and ZZ. This can be done by observing the appropriate measurements (which form a probability distribution over four outcomes) and applying a Walsh-Hadamard transform on that probability distribution. The Walsh-Hadamard transform is a form of Fourier transform that moves from a probability distribution (which may be sparse) to the dense eigenvalue basis of the channel and vice versa (for details, again see Ref. [48]).

APPENDIX F: PARAMETER ESTIMATION FOR THE FACTORS

Here, we give the intuition behind the scalable creation of a Markov network such as the Ising model using the locally averaged noise distribution. As mentioned previously, for a device with n data qubits, this is just a classical probability distribution of size 2^n. Consider such a distribution over n bits $\mathbf{x} = x_1 x_2 \ldots x_n$, which we can write as $p(x)$. We are looking to efficiently represent this distribution. One well-known approach is to model it as an undirected graphical model or Gibbs random field [79]. Here, the probability distribution is associated with variables that live on the "variable" nodes of a factor graph and the interactions live on the "factor" nodes that describe how to couple the variables, as we now detail.

The underlying assumption behind such a graphical model is the *global Markov* property, which is contained in the structure of each particular factor graph G. In such a graph [see, e.g., Fig. 9(a)], we associate each variable x_j with a data qubit, which we also label x_j. The "factor" nodes connect these variables and are labeled by the factors, here ψ_i. These describe the correlations that are possible in p according to the *global Markov property*.

To define the global Markov property, let us assume that there is a separating set S of vertices with variables $\mathbf{x}_S$ between two sets of variables $\mathbf{x}_A$ and $\mathbf{x}_B$ (i.e., every path in G between A and B goes through at least one vertex in S). Then, the global Markov property asserts that

$$p(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_S) = p(\mathbf{x}_A|\mathbf{x}_S)p(\mathbf{x}_B|\mathbf{x}_S) \tag{F1}$$

for every A and B and any separating set S. That is, that the marginal distribution over x_A and x_B is conditionally independent given the values on the separating subset $\mathbf{x}_S$. We illustrate this with examples in Fig. 9.

The Hammersley-Clifford theorem [79] says that every strictly positive probability distribution that obeys the global Markov property for a factor graph G factorizes over the factor nodes of G, such as the graph shown in Fig. 9(a). That is, $p(\mathbf{x}) = 1/Z \prod_k \psi_k$, where ψ_k are the factor functions, Z is a normalization, and the factors ψ_k for each factor node are positive functions.

Because the functions ψ_k are positive, they can be (and often are) reparametrized as $\psi_k = \exp(H_k)$, where H_k is a real-valued function (an abstract "Hamiltonian") that is a function only of the variables in a neighborhood of at most one factor. In that case, the normalizing factor Z takes on the interpretation of the partition function of the Gibbs distribution of the total Hamiltonian $H = \sum_k H_k$ at an inverse temperature $\beta = 1$. This physics language is helpful but we gently remind the Ising-model aficionados that the variables here are bits, $x \in \{0, 1\}$, so some intuitions coming from spin variables can be misleading in this context.

A particularly convenient form for the Hamiltonian H is to express it as a sum of products of the variables nodes via the following construction. For every factor k in the factor graph, we can replace the factor node and its neighborhood by the local complement in the graph. That is, all the variable nodes connected to the factor k are connected into a complete graph (since they were previously nonadjacent) and the factor is disconnected from the graph. We call this modified factor graph (with the disconnected factor nodes now discarded) G'. Finally, let C be the set of all cliques in G', i.e., the set of all complete induced subgraphs. We

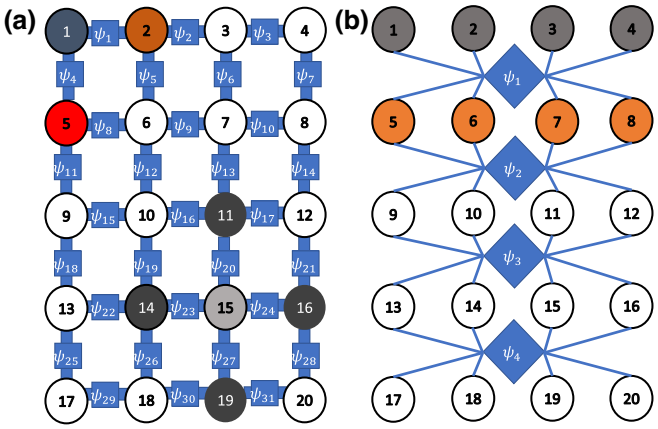

FIG. 9. An illustration of the Markov-blanket assumption built into the Ising-model factor graph and the CG1D graph. As discussed in the text, the errors on each qubit are independent of all other qubits if one is given the qubits that share a factor with the qubit in question. The variables are the data qubits and are associated with the circular "variable" nodes in the graphs, labeled 1 through 20. The square (diagonal) nodes represent the "factor" nodes. On the graphs, we have color coded some of the qubits to aid with identification in the discussion below. (a) Qubit 1 (dark gray) has two factors (ψ_1 and ψ_3) that touch it. The only other qubits that these factors touch are qubits 2 and 5 (red). Therefore, the assumption is that the errors on qubit 1 are independent of the other qubits (excluding qubits 2 and 5), given the errors on qubits 2 and 5. Similarly, the Markov blanket on qubit 15 (light gray) is formed from qubits 11, 14, 16, and 19 (black). This means that the assumption built into the model is that the distribution on qubit 15 is independent of the other qubits, if one knows the errors on qubits 11, 14, 16, and 19. There are 31 factors required in this graph and each factor specifies a 2^2 distribution, meaning that the model requires 124 parameters. (b) This model is referred to in the text as the CG1D model, where we have reduced the surface code to a 1D graph. In this case, the independence assumptions can be read from top to bottom and we have that qubits 1, 2, 3, and 4 (light gray) are independent of qubits 9 through to 20, given qubits 5, 6, 7, and 8 (orange). There are only four factors in this model, but each factor has a distribution of 2^8, meaning that this model requires 1024 parameters to be fully specified. This is the only model with exponential scaling; it scales exponentially in the width of the grid. We note that the choice between width and height factorization is arbitrary; in this case, we have chosen the width because the factors would be smaller. In practice, it makes little difference to the results presented.

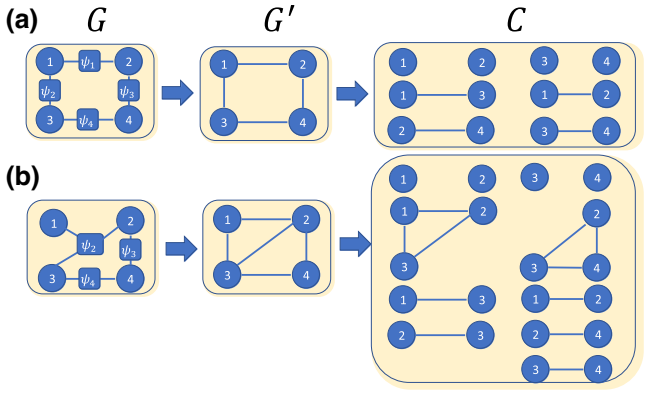

FIG. 10. An illustration of the conversion of a factor graph into the set of relevant cliques for two different types of factor graph. In the first step, each node touching a factor is connected to the nodes also touching that factor and the factor is removed. In the second step, we create a set of all of the induced subgraphs.

illustrate the transformation from G to G' and the set of cliques C in Fig. 10. With this notation, we have, for some real numbers J_b, the general form

$$H(\mathbf{x}) = -\sum_{b\in C} J_b \prod_{a\in b} x_a. \tag{F2}$$

The minus sign in front is purely a convention. Each of these terms functions as a local "coupling" that is local to the clique b and that lowers the energy by J_b whenever all variables in b are equal to 1.

We can also see from this mapping that $H(\mathbf{0}) = -J_\emptyset$ is a constant "energy shift." Having already defined a partition function, this factor is redundant and could be set to zero. With that convention, we have the result that $p(\mathbf{0}) = 1/Z\prod_k \exp(H(\mathbf{0})) = 1/Z$. Alternatively, we could omit the partition function (setting $Z = 1$) and keep $J_\emptyset = -\log p(\mathbf{0})$ as the "free energy" of the Gibbs ensemble. This second convention is sometimes useful for bookkeeping in calculations. Either way, this normalizing factor is likely hard to compute in general given the J_b. However, for the very small error rates in quantum computers, it might be possible to obtain a direct estimate of $p(\mathbf{0})$ from sampled data. Unfortunately, as the system size becomes large enough, this eventually becomes exponentially small, and therefore impractical to estimate through sampling.

Because of this, we relax our goal to specifying the entire probability distribution $p(\mathbf{x})$ *up to normalization* and this is now equivalent to specifying the values of the coupling constants J_b, one for each nonempty clique b in G'. If there are at most f factors with neighborhoods in G of size at most d, then the entire probability distribution is specified by at most $f\,2^d$ real numbers, which is much less than 2^n in the regime of interest, where f is linear in n and d is constant.

How might one instantiate these factors, i.e., learn the parameters J_b given knowledge (or an assumption) about the factor graph G? First, consider a set of variables $\mathbf{x}_r$, where $r \in C$. Suppose that we consider an event where every variable outside of r, denoted $\mathbf{x}_{r^c}$, is zero. Then, any product of x_a that involves a variable outside of r vanishes. That is, if $b \not\subseteq r$, then $\prod_{a\in b} x_a$ contains at least one 0. The log probability then greatly simplifies to

$$-\log p(\mathbf{x}_r, \mathbf{0}_{r^c}) = \sum_{b\in C} J_b \prod_{a\in b} x_a \Big|_{\mathbf{x}_{r^c}=\mathbf{0}} \tag{F3}$$

$$= \sum_{b\subseteq r} J_b \prod_{a\in b} x_a. \tag{F4}$$

This gives a set of linear equations that relate the log probabilities with the coupling constants when we enumerate over the $2^{|r|}$ bit strings $\mathbf{x}_r$. To illustrate, suppose that $|r| = 2$, and write $p(x_1x_2, \mathbf{0})$ for $p(\mathbf{x}_r, \mathbf{0}_{r^c})$. Then, we would have the four equations

$$\begin{aligned}
-\log p(00,\mathbf{0}) &= J_\emptyset, \\
-\log p(01,\mathbf{0}) &= J_\emptyset + J_{01}, \\
-\log p(10,\mathbf{0}) &= J_\emptyset + J_{10}, \\
-\log p(11,\mathbf{0}) &= J_\emptyset + J_{01} + J_{10} + J_{11}.
\end{aligned} \tag{F5}$$

Here, we have adopted the convention that the label for the subset b is just the bit string with 1 as the indicator function for set membership.

We now make two important simplifications to this problem. The probabilities $p(\mathbf{x}_r, \mathbf{0}_{r^c})$ will be extremely small in general, even for independent noise. These cannot be efficiently learned from sampling, even if $|r|$ is a small constant, since they decay exponentially in n, the total number of bits. We therefore note that we can use our relaxed goal of learning modulo the normalization to add $\log p(\mathbf{0}_{r_c})$ to both sides. By Bayes' theorem, we have $p(\mathbf{x}_r, \mathbf{0}_{r_c}) = p(\mathbf{x}_r|\mathbf{0}_{r_c})p(\mathbf{0}_{r_c})$, so the left-hand side becomes a conditional probability, $-\log p(\mathbf{x}_r|\mathbf{0}_{r_c})$. Let us write ∂r for the set of nodes in the neighborhood of r. Then, ∂r forms a separating set and using the global Markov property implies that the left-hand side simplifies further to $-\log p(\mathbf{x}_r|\mathbf{0}_{\partial r})$. This boundary set ∂r will have bounded size if each of the factors have bounded size and if each variable participates in a bounded number of factors. Thus, the left-hand side becomes conditional probability distributions over a bounded number of variables, for which an empirical estimation can be done with reasonable efficiency. To balance the right-hand side, the term $J_\emptyset = -\log p(\mathbf{0})$ transforms to $J_{\mathbf{0}_r} := -\log p(\mathbf{0}_r|\mathbf{0}_{\partial r})$. Note that $J_{\mathbf{0}_r}$ does not appear anywhere in our Hamiltonian [Eq. (F2)], so our estimate of this is in some sense a nuisance parameter.

To summarize, we now have an equivalent system of equations where the quantities on the left-hand side allow

reasonable empirical estimates, given in illustration for $|r| = 2$ by

$$\begin{array}{lcllll} -\log p(00|\mathbf{0}_{\partial r}) & = & J_{00}, & & & \\ -\log p(01|\mathbf{0}_{\partial r}) & = & J_{00} & +J_{01}, & & \\ -\log p(10|\mathbf{0}_{\partial r}) & = & J_{00} & & +J_{10}, & \\ -\log p(11|\mathbf{0}_{\partial r}) & = & J_{00} & +J_{01} & +J_{10} & +J_{11}. \end{array} \tag{F6}$$

We have only to solve these equations for the $J_{\mathbf{x}_r}$ for $\mathbf{x}_r \neq \mathbf{0}_r$.

With this binary ordering, we define a matrix

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{\otimes |r|}, \tag{F7}$$

so that in general our equations have the simple form

$$-\log p(\mathbf{x}_r|\mathbf{0}_{\partial r}) = AJ_{\mathbf{x}_r}. \tag{F8}$$

Note that A is invertible, with inverse

$$A^{-1} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}^{\otimes |r|}. \tag{F9}$$

Therefore, a general solution for the couplings in the subset r is given by

$$J_{\mathbf{x}_r} = -A^{-1}\big(\log p(\mathbf{x}_r|\mathbf{0}_{\partial r})\big). \tag{F10}$$

By replacing the conditional probabilities by empirical estimates of conditional probabilities, we directly obtain empirical estimates for the coupling constants in any given clique r.

We need to comment on the theoretical precision of this estimation procedure. First, the condition number of A is $\big(3+\sqrt{5}/2\big)^{|r|} \approx 2.61^{|r|}$, so the linear inversion becomes poorly conditioned for large $|r|$. Second, a useful estimate of $-\log p$ requires an estimate of p to *relative* precision $p(1 \pm \epsilon)$ in order for the error to remain tolerable after the logarithm. This is expensive when p is small. Despite these challenging theoretical limitations, we empirically observe quite favorable performance of this procedure. We leave the search for even better estimators with improved guarantees open to future work.

The estimation problem substantially simplifies in the case where the factor graph is 1D or, as in the case of Fig. 9(b), is 1D up to some coarse graining. In that case, one only needs to estimate the marginals (as opposed to conditional marginals) on nearest neighbors along the 1D geometry. This follows from a repeated application of Bayes' theorem together with the global Markov property, as we now show. Consider a chain of variables $\mathbf{x}_{1:n} = x_1 x_2 \ldots x_n$, which need not be binary as they may have arisen from some coarse graining. By applying Bayes' theorem followed by the global Markov property, we have

$$\begin{aligned} p(\mathbf{x}_{1:n}) &= p(x_1|\mathbf{x}_{2:n})p(\mathbf{x}_{2:n}) \\ &= p(x_1|x_2)p(\mathbf{x}_{2:n}) \\ &= \frac{p(x_1,x_2)}{p(x_2)}p(\mathbf{x}_{2:n}). \end{aligned}$$

We can now recursively apply this idea to the rest of the chain starting at x_2 and we find

$$p(\mathbf{x}_{1:n}) = \prod_{k=1}^{n-1} \frac{p(x_k, x_{k+1})}{p(x_{k+1})} \times p(x_n).$$

In this way, the case of a 1D structure allows some additional efficiency in estimating the model parameters, since they can now be reconstructed from (estimates of) only nearest-neighbor marginals.

In the case of the type of model proposed for the surface code, calculation of any particular factor will involve a maximum of eight qubits. As previously discussed, for locally averaged noise, the joint distributions for all possible eight-qubit groupings can be ascertained with a single "RB-style" experiment. While this calculation is predicated on the fact that the underlying probability distribution does indeed obey the global Markov global property (which will not necessarily be the case for the actual noise in the device), Ref. [80] shows (and our numerics indicate in an experimental setting) that the process degrades gracefully, i.e., the calculated approximation will indeed approximate the underlying probability distribution. Note that although we can populate the factors, calculating the partition function (needed to normalize the graph so that it forms a probability distribution), still requires a calculation that scales exponentially. This is, however, not a problem when we do not need to know the probabilities but just the relative probabilities. Importantly, this means that we can sample from the model in a scalable fashion in most cases of interest.

APPENDIX G: COVARIANCE BOUNDS

Suppose that we have two finite probability distributions, p and q. Now consider a random variable X, which is a real vector that is distributed according to either p or q. Let the covariance matrix of X and the mean of X with respect to the distribution p be

$$\begin{aligned} \Sigma_p &= \mathbb{E}_p\big[(X-\mu_p)(X-\mu_p)^T\big] \\ &= \mathbb{E}_p\big[XX^T\big] - \mu_p\mu_p^T, \end{aligned} \tag{G1}$$

$$\mu_p = \mathbb{E}_p[X]. \tag{G2}$$

We would like to prove an upper bound on the operator norm of the difference, $\|\Sigma_p - \Sigma_q\|$, in terms of a distance

between p and q and some notion of the length of the X_j. We have the following theorem.

Theorem 1.—Let $C = \mathrm{conv}\{X_j | j \in S\}$ be the convex hull of the vectors X_j in $S = \mathrm{supp}(p-q)$, the support of $p-q$. Let $D = \mathrm{diam}(C)$ be the diameter of C in the Euclidean distance and let $T = T(p,q) = \frac{1}{2}\|p-q\|_1$ be the statistical (1-norm) distance between p and q. Then, we have

$$\|\Sigma_p - \Sigma_q\| \leq T\big(D^2 - \tfrac{1}{4}\|\mu_p - \mu_q\|^2\big). \tag{G3}$$

Moreover, the scaling TD^2 cannot be sharpened by any constant factor.

Proof.—We introduce some new variables,

$$m = \frac{p+q}{2}, \tag{G4}$$

$$\mu = \mathbb{E}_m[X] = \frac{\mu_p + \mu_q}{2}, \tag{G5}$$

$$\delta = \frac{\mu_p - \mu_q}{2}, \tag{G6}$$

$$Y_j = X_j - \mu, \tag{G7}$$

so that $\mu_p = \mu + \delta$, $\mu_q = \mu - \delta$, and $\mathbb{E}_m[Y] = 0$. In terms of these variables, we have

$$\begin{aligned}
&\Sigma_p - \Sigma_q \\
&\quad = \mathbb{E}_p\big[(Y-\delta)(Y-\delta)^T\big] - \mathbb{E}_q\big[(Y+\delta)(Y+\delta)^T\big] \\
&\quad = \sum_j (p_j - q_j) Y_j Y_j^T - \sum_j (p_j + q_j)(\delta Y_j^T + Y_j \delta^T) \\
&\quad + \sum_j (p_j - q_j)\delta\delta^T \\
&\quad = \sum_j (p_j - q_j) Y_j Y_j^T.
\end{aligned} \tag{G8}$$

In the last step, the middle terms vanish because $\mathbb{E}_m[Y] = 0$ and the last term vanishes identically because $\sum_j (p_j - q_j) = 0$. Now, we split the sum into two sets, given by

$$j_+ = \{j : p_j > q_j\}, \quad j_- = \{j : p_j < q_j\}, \tag{G9}$$

so that taking norms of both sides, we have

$$\begin{aligned}
&\|\Sigma_p - \Sigma_q\| \\
&\quad = \left\| \sum_j (p_j - q_j) Y_j Y_j^T \right\| \\
&\quad = \left\| \sum_{j\in j_+} |p_j - q_j| Y_j Y_j^T - \sum_{j\in j_-} |p_j - q_j| Y_j Y_j^T \right\| \\
&\quad = \|M_+ - M_-\|.
\end{aligned} \tag{G10}$$

In the last equality, the matrices $M_\pm$ that we introduce are both positive semidefinite, $M_\pm \geq 0$, since they are positive sums of positive semidefinite matrices. But for any two positive semidefinite matrices $M_\pm$, we have

$$\begin{aligned}
&\|M_+ - M_-\| \\
&\quad = \max_{v:\|v\|=1} |v^T(M_+ - M_-)v| \leq \max\{\|M_+\|, \|M_-\|\}.
\end{aligned} \tag{G11}$$

We will assume without loss of generality that $\|M_+\| \geq \|M_-\|$. Because both p and q are probability distributions, we have

$$\sum_{j_+} |p_j - q_j| = \sum_{j_-} |p_j - q_j| = \frac{1}{2}\|p-q\|_1 = T(p,q). \tag{G12}$$

Using the inequality from Eq. (G11), the triangle inequality, Hölder's inequality, and Eq. (G12), we have

$$\begin{aligned}
\|\Sigma_p - \Sigma_q\| &\leq \max\{\|M_+\|, \|M_-\|\} && \text{(G13)} \\
&\leq \sum_{j\in j_+} |p_j - q_j| \left\| Y_j Y_j^T \right\| && \text{(G14)} \\
&\leq \sum_{j\in j_+} |p_j - q_j| \max_{k\in S} \left\| Y_k Y_k^T \right\| && \text{(G15)} \\
&= T \max_{k\in S} \left\| X_k - \mu \right\|^2. && \text{(G16)}
\end{aligned}$$

Geometrically, $\max_{k\in S}\|X_k - \mu\|$ is the maximum distance between the vectors X_k and some point μ that must be inside C, their convex hull. This distance is clearly bounded by D, the diameter of C, and leads to the weaker bound of $\|\Sigma_p - \Sigma_q\| \leq D^2 T$.

To obtain the sharper statement of the theorem, we need one more observation. Define $X^\star = X_{k^\star}$, where $k^\star = \arg\max_{k\in S}\|X_k - \mu\|$ is the index of any maximizer (if there is more than one, we choose one arbitrarily). Because the diameter of C is D, we must also have $r_p := \|X^\star - \mu_p\| \leq D$ and $r_q := \|X^\star - \mu_q\| \leq D$. But the distance $\|\mu_p - \mu_q\|$ is fixed, so the distance of interest, $r := \|X^\star - \mu\|$, cannot be as large as D unless $p = q$, as the distance to the midpoint will generally be shorter. The maximizing distance will, in fact, be the median of the triangle formed by the three points μ_p, μ_q, and $X^\star$. Elementary geometry shows that this squared distance is then bounded as follows

$$\begin{aligned}
r^2 &= \|X^\star - \mu\|^2 \\
&= \frac{2r_p^2 + 2r_q^2 - \|\mu_p - \mu_q\|^2}{4} \leq D^2 - \frac{\|\mu_p - \mu_q\|^2}{4}.
\end{aligned} \tag{G17}$$

This completes the proof of the stronger inequality.

To show that the upper bound cannot be strengthened to $c\,TD^2$ for any constant $c < 1$, consider the following

example. For the probability distributions $p = (1, 0)$, $q = (1 - a, a)$ (with $1 \geq a \geq 0$), we have

$$T = \tfrac{1}{2}\|p - q\|_1 = a. \tag{G18}$$

If we consider the scalar random variables $X_0 = 0$, $X_1 = D$, then the covariance with respect to each probability distribution is just the variance. Only q yields a nontrivial variance given by $D^2 a(1 - a)$, so the bound is indeed tight to leading order in a. ■

In the case of reconstructing the $(n \times n)$ Pauli covariance matrix, we have guaranteed convergence in the 1-norm between p and an estimate q of p and the vectors X_j have diameter $D = \sqrt{n}$, where n is the number of qubits. So the covariance estimate that we report using the initial density estimate is consistent, though it may have some bias. Moreover, the scaling of the upper bound proven in the theorem is most likely an artifact in that case, for the following reason. Both p_j and q_j get smaller when j labels higher-weight errors, so $\|X_j\|$ gets suppressed for those values in the sum. The bound is worst case and is not sensitive to this. It only takes a $1/w$ dependence with the weight to remove the scaling with n, so it is very likely that the n dependence is not there in practice.

Finally, we note it is not possible to upper bound the total variation distance between two finite probability distributions on bit strings by a function only of the covariance matrix of the random variable.

Consider q being the uniform distribution on all strings of length $n = 2^k$ and p being uniform on the $2n$ strings that come from the rows of a Hadamard matrix and its complement. The TVD between p and q is easily computed to be $1 - n2^{1-n}$. But the covariance between p and q is identically 0. In fact, the first and second moments are both zero.

APPENDIX H: JUSTIFYING THE ISING MODEL WITH DATA

As noted in the main text, for this specific device, the Ising model is unable to capture some of the longer-range correlated errors that appear to exist in the device. This begs the question: is there another type of two-factor graph that we can draw (say, by way of example) connecting qubits 1 and 9 rather than 1 and 2) that might better represent the underlying probability distribution?

Given that we have access to the full distribution this is, in fact, a question we can attempt to answer.

For (classical) random variables $\mathcal{X}$ and $\mathcal{Y}$, the *conditional entropy* (CE) of $\mathcal{Y}$ given $\mathcal{X}$ is defined as

$$H(\mathcal{Y}|\mathcal{X}) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{H1}$$

and quantifies the amount of information needed to describe the outcome of random variable $\mathcal{Y}$ given that we have knowledge of the value of $\mathcal{X}$. We can use this measure to determine the optimal qubits to link together in our graph. The lower the conditional entropy of $\mathcal{Y}$, the "stronger" the link to $\mathcal{X}$. If we specify a maximum number of links (n) which are allowed to any particular qubit (q). Then, for each candidate qubit (q), we can search to see which group of n qubits will give us the lowest conditional entropy. Those are the qubits we might consider linking in our graph.

If we decide to limit the number of "linked" qubits to be the same as in the Ising model (e.g., for a corner qubit we want to draw a link to two other qubits, for a middle qubit we will allow it to link to four other qubits), then a brute-force search for the best qubits to link together is quite tractable in the current device.

Carrying out this procedure, we find that the extracted data confirm that the Ising-model links are the strongest links with three exceptions, which are set out in Table II. Even in those exceptional cases, the "optimal" links are extremely close to the Ising-model links, differing in all cases by only one qubit and a conditional entropy of less than 0.5% from the Ising-ansatz conditional entropy. This process was repeated 100 times with different distributions, bootstrapped from the original data (which gives us the confidence level described in the table). The results were consistent, providing a large degree of confidence that the Ising-model ansatz based on qubit physical location is a sensible ansatz to use.

TABLE II. For each of the data qubits in the device (labeled in the order shown in Fig. 4), the qubits that are linked to the qubit in question under the Ising model have been read from the graph. All possible combinations of the same number of data qubits have been assessed using conditional entropy (CE) and the qubits with the lowest value have been found. Other than as listed above, they are the same as for the Ising-model ansatz. We note that even in those cases, all but one of the qubits in each "optimal" blanket are the same as in the Ising model and that the difference in the CE between the optimal qubits and the Ising ansatz is, in all cases, less than 0.5%. The use of a bootstrap methodology to ascertain error bars shows us that the 95% confidence level in all cases is less than $\pm 2 \times 10^{-5}$.

Data qubit	Alternative qubits	CE alternative	CE Ising	Ising qubits
6	[5, 7, 10, 11]	0.8166	0.8167	[5, 7, 2, 10]
13	[9, 14, 18]	0.874	0.875	[9, 14, 17]
18	[13, 14, 17]	0.797	0.799	[14, 17, 19]

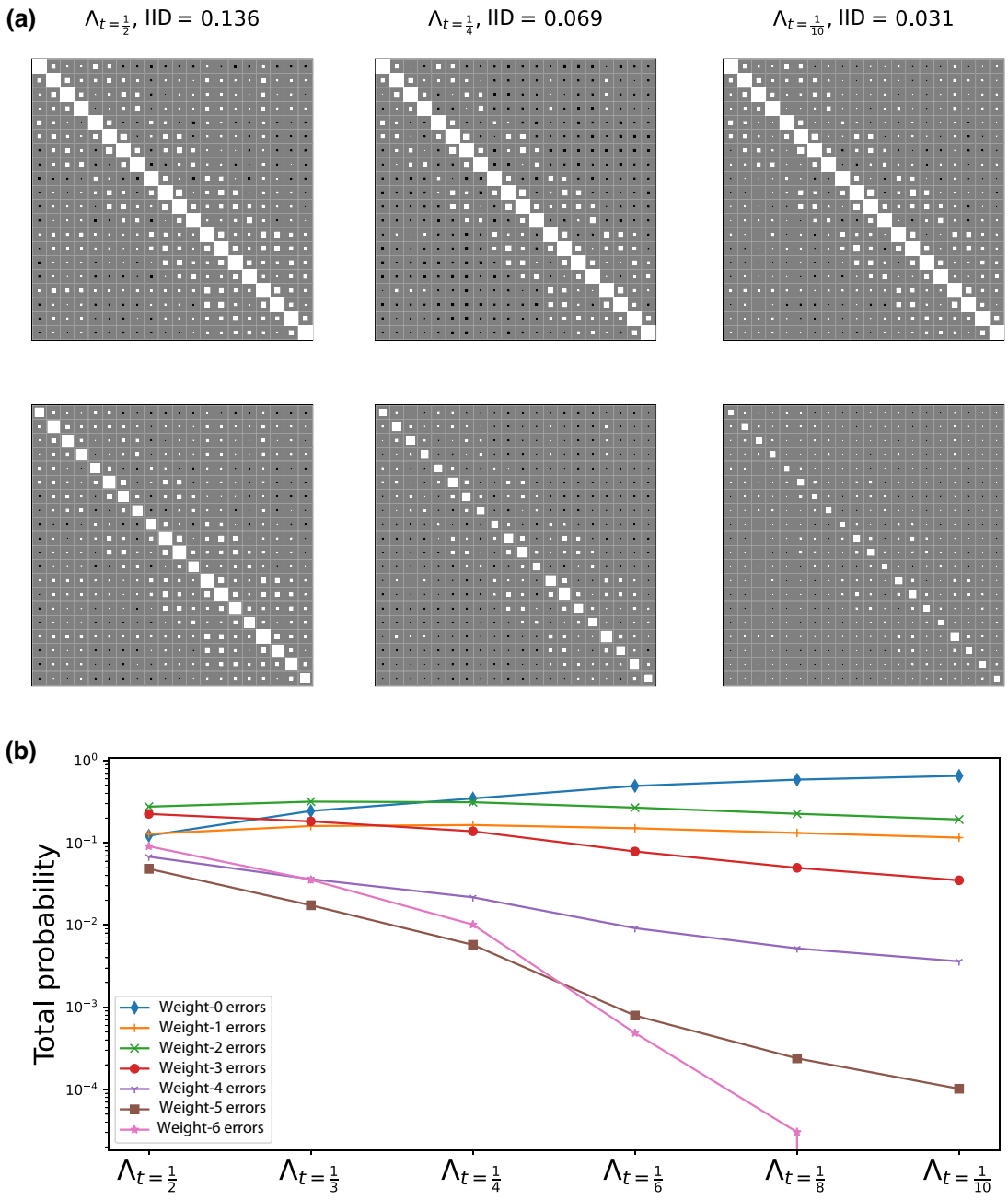

FIG. 11. (a) Hinton plots for the correlation (top row) and covariance (second row) matrices for some example reconstructed noise channels (see Appendix I). The qubits are set out as data qubit 0 → data qubit 19. The value of the correlation and/or covariance is related to the area of the square. For the correlation matrices, a full square has the value of 1. For instance, all squares on a diagonal are 1 (a full white square), as each qubit is correlated with itself. White represents positive correlations, black negative ones. For the covariance matrices, the scale is kept constant at a value of a full square = 0.18 (which is the largest value in all three example plots). The first example, $\Lambda_{t=\frac{1}{2}}$, shows the correlation matrix for the noise channel representing a single round of the stabilizer-preparation circuits. The title for each diagram also shows the average of the marginalized data qubit error rates (the number used in the simple IID model). The other diagrams show the same data for different values of t, e.g., the $\Lambda_{t=1/4}$ diagram represents the correlation matrix for the reconstructed channel that, if applied twice, would result in the same channel as $\Lambda_{t=1/2}$. As can be seen, while the average qubit error rate decreases as $t \to 0$, the correlation pattern of each channel remains reasonably consistent and, importantly, retains the interesting noise features of the device. As might be expected, the raw value of the covariance between the qubits decreases as the fidelity of the channel increases but the patterns remain consistent. (b) The total probability for errors with specific "weights," i.e., affecting the specified number of qubits. The plot shows the prevalence of high-weight errors with different noise maps. As can be seen, as t decreases, there are fewer high-weight errors. This is as anticipated, since as these higher-fidelity maps are applied multiple times, lower-weight errors can combine to form higher-weight errors. The noise on the device is such, though, that errors with weights ≥ 3 are still prevalent even on the least noisy maps.

APPENDIX I: RECONSTRUCTED NOISE CHANNELS

As discussed in the main text, the Walsh-Hadamard transform can be used to move between the observed error-probability domain and the Pauli channel eigenvalues (or, as here, the locally averaged channel eigenvalues). Using this transform, it is trivial to calculate the effect of multiple applications of the channel, even when one starts from an observed probability distribution. For instance, two applications of the channel are calculated by squaring the eigenvalues. It is exactly this property that is used in fitting decay curves in randomized benchmarking.

Similarly, one can easily model the effect of "partial" applications of the channel; e.g., by taking the square root of the eigenvalues, one can calculate a channel that if applied twice would result in the original observed channel. This is most easily seen if one recalls that a superoperator representation of a Pauli channel (in a Pauli basis) is a diagonal matrix, with the diagonal elements being the Pauli eigenvalues of the channel. Taking the square root of these diagonal elements results in a channel that when applied twice (multiplied by itself) results in the original channel.

We can make this more precise if one assumes that the noise is generated by a continuous-time process. In that case, let p be the probability distribution measured by our protocol and let W be the Walsh-Hadamard transform. Then, combining Eqs. (3) and (4) we have

$$p(t) = W^{-1}\exp(t\log(Wp)). \tag{I1}$$

As noted in the main text, not every noise channel, including Pauli noise channels, is divisible in this fashion. The issue is that the object constructed in the manner of Eq. (I1) might not correspond to a completely positive trace-preserving map. Indeed, in practice, especially given the size of the channels and numeric imprecision, the raw transformed channel tended to have a number of small negative "probabilities." As a practical matter, we then took the step of projecting the resulting distributions onto the nearest probability simplex. It should be restated that the purpose of this exercise is not to generate channels that represented something *achievable* in the device but, rather, to "construct" counterfactual theoretical channels that have lower error rates than the observed distribution but that also, so far as possible, retain all the "interesting" features and correlations of the original measured noise. For instance, Fig. 11(a) shows that these theoretical channels contain two-body correlations that appear broadly similar to the original observed channel and to each other.

Finally, we would anticipate that the higher-fidelity noise maps (which represent partial applications of the observed noise map) would have fewer high-weight errors as a percentage of the total probability "budget." This is because as they are applied multiple times, smaller-weight errors can combine into larger-weight errors, e.g., an IX error and XI error will combine to an XX error. Figure 11(b) shows the prevalence of errors by weight in the various noise maps, confirming this behavior.

Where our observed channel might not be below threshold (i.e., the logical error rate is above the physical error rate), this will allow one to explore where (or if) the threshold is crossed, given the *characteristics* of the noise.

[1] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, SIAM J. Comput. **26**, 1484 (1997).

[2] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik, Quantum chemistry in the age of quantum computing, Chem. Rev. **119**, 10856 (2019).

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

[4] A. Aspuru-Guzik, A. D. Dutoi, P. J. Love, and M. Head-Gordon, Simulated quantum computation of molecular energies, Science **309**, 1704 (2005).

[5] C. Gidney and M. Ekerå, How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits, Quantum **5**, 433 (2021).

[6] N. Grzesiak, R. Blümel, K. Wright, K. M. Beck, N. C. Pisenti, M. Li, V. Chaplin, J. M. Amini, S. Debnath, J.-S. Chen, and Y. Nam, Efficient arbitrary simultaneously entangling gates on a trapped-ion quantum computer, Nat. Commun. **11**, 2963 (2020).

[7] W. Huang, C. H. Yang, K. W. Chan, T. Tanttu, B. Hensen, R. C. C. Leon, M. A. Fogarty, J. C. C. Hwang, F. E. Hudson, K. M. Itoh, A. Morello, A. Laucht, and A. S. Dzurak, Fidelity benchmarks for two-qubit gates in silicon, Nature **569**, 532 (2019).

[8] E. H. Chen, T. J. Yoder, Y. Kim, N. Sundaresan, S. Srinivasan, M. Li, A. D. Córcoles, A. W. Cross, and M. Takita, Calibrated decoders for experimental quantum error correction, Phys. Rev. Lett. **128**, 110504 (2022).

[9] C. J. Ballance, T. P. Harty, N. M. Linke, M. A. Sepiol, and D. M. Lucas, High-fidelity quantum logic gates using trapped-ion hyperfine qubits, Phys. Rev. Lett. **117**, 060504 (2016).

[10] E. Knill, R. Laflamme, and W. H. Zurek, Resilient quantum computation, Science **279**, 342 (1998).

[11] D. Aharonov and M. Ben-Or, in *29th ACM Symp. on Theory of Computing (STOC)* (New York, USA, 1997), p. 176.

[12] A. Y. Kitaev, Quantum computations: Algorithms and error correction, Russ. Math. Surv. **52**, 1191 (1997).

[13] C. Ryan-Anderson, J. G. Bohnet, K. Lee, D. Gresh, A. Hankin, J. P. Gaebler, D. Francois, A. Chernoguzov, D. Lucchetti, N. C. Brown, T. M. Gatterman, S. K. Halit, K. Gilmore, J. A. Gerber, B. Neyenhuis, D. Hayes, and R. P. Stutz, Realization of real-time fault-tolerant quantum error correction, Phys. Rev. X **11**, 041058 (2021).

[14] L. Egan, D. M. Debroy, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Newman, M. Li, K. R. Brown, M. Cetina, and C. Monroe, Fault-tolerant control of an error-corrected qubit, Nature **598**, 281 (2021).
[15] S. Krinner, N. Lacroix, A. Remm, A. D. Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, Realizing repeated quantum error correction in a distance-three surface code, Nature **605**, 669 (2022).
[16] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, Matching and maximum likelihood decoding of a multi-round subsystem quantum error correction experiment (2022).
[17] Y. Zhao, *et al.*, Realization of an error-correcting surface code with superconducting qubits, Phys. Rev. Lett. **129**, 030501 (2022).
[18] A. Kitaev, Fault-tolerant quantum computation by anyons, Ann. Phys. (NY) **303**, 2 (2003).
[19] S. B. Bravyi and A. Y. Kitaev, Quantum codes on a lattice with boundary (1998).
[20] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Surface codes: Towards practical large-scale quantum computation, Phys. Rev. A **86**, 032324 (2012).
[21] D. K. Tuckett, S. D. Bartlett, and S. T. Flammia, Ultrahigh error threshold for surface codes with biased noise, Phys. Rev. Lett. **120**, 050505 (2018).
[22] J. P. Bonilla Ataides, D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, The XZZX surface code, Nat. Commun. **12**, 2172 (2021).
[23] R. Acharya, *et al.*, Suppressing quantum errors by scaling a surface code logical qubit, Nature **614**, 676 (2023).
[24] A. Robertson, C. Granade, S. D. Bartlett, and S. T. Flammia, Tailored codes for small quantum memories, Phys. Rev. Appl. **8**, 064004 (2017).
[25] N. H. Nickerson and B. J. Brown, Analysing correlated noise on the surface code using adaptive algorithms, Quantum **3**, 131 (2019).
[26] D. K. Tuckett, A. S. Darmawan, C. T. Chubb, S. Bravyi, S. D. Bartlett, and S. T. Flammia, Tailoring surface codes for highly biased noise, Phys. Rev. X **9**, 041031 (2019).
[27] C. T. Chubb, General tensor network decoding of 2D Pauli codes (2021).
[28] C. T. Chubb and S. T. Flammia, Statistical mechanical models for quantum codes with correlated noise, Ann. de l'Inst. Henri Poincaré D **8**, 269 (2021).
[29] D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, Fault-tolerant thresholds for the sufrace code in excess of 5% under biased noise, Phys. Rev. Lett. **124**, 130501 (2020).
[30] A. deM. iOlius, J. E. Martinez, P. Fuentes, P. M. Crespo, and J. Garcia-Frias, Performance of surface codes in realistic quantum hardware (2022).
[31] K. Tiurev, P.-J. H. S. Derks, J. Roffe, J. Eisert, and J.-M. Reiner, Correcting non-independent and non-identically distributed errors with surface codes (2022).
[32] O. Higgott, T. C. Bohdanowicz, A. Kubica, S. T. Flammia, and E. T. Campbell, Fragile boundaries of tailored surface codes and improved decoding of circuit-level noise (2022).
[33] 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).
[34] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Self-consistent quantum process tomography, Phys. Rev. A **87**, 062119 (2013).
[35] R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz, Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography, Nat. Commun. **8**, 14485 (2016).
[36] T. Evans, W. Huang, J. Yoneda, R. Harper, T. Tanttu, K. Chan, F. Hudson, K. Itoh, A. Saraiva, C. Yang, A. Dzurak, and S. Bartlett, Fast Bayesian tomography of a two-qubit gate set in silicon, Phys. Rev. Appl. **17**, 024068 (2022).
[37] J. Emerson, M. Silva, O. Moussa, C. Ryan, M. Laforest, J. Baugh, D. G. Cory, and R. Laflamme, Symmetrized characterization of noisy quantum processes, Science **317**, 1893 (2007).
[38] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).
[39] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and robust randomized benchmarking of quantum processes, Phys. Rev. Lett. **106**, 180504 (2011).
[40] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, Characterizing universal gate sets via dihedral benchmarking, Phys. Rev. A **92**, 060302 (2015).
[41] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, Real randomized benchmarking, Quantum **2**, 85 (2018).
[42] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, Characterizing large-scale quantum computers via cycle benchmarking, Nat. Commun. **10**, 5347 (2019).
[43] J. Helsen, X. Xue, L. M. Vandersypen, and S. Wehner, A new class of efficient randomized benchmarking protocols, npj Quantum Inf. **5**, 71 (2019).
[44] T. J. Proctor, A. Carignan-Dugas, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, Direct randomized benchmarking for multiqubit devices, Phys. Rev. Lett. **123**, 030503 (2019).
[45] T. Proctor, S. Seritan, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, Scalable randomized benchmarking of quantum computers using mirror circuits, Phys. Rev. Lett. **129**, 150502 (2022).
[46] J. Hines, M. Lu, R. K. Naik, A. Hashim, J.-L. Ville, B. Mitchell, J. M. Kriekebaum, D. I. Santiago, S. Seritan, E. Nielsen, R. Blume-Kohout, K. Young, I. Siddiqi, B. Whaley, and T. Proctor, Demonstrating scalable randomized benchmarking of universal gate sets (2022).
[47] T. Proctor, K. Rudinger, K. Young, E. Nielsen, and R. Blume-Kohout, Measuring the capabilities of quantum computers, Nat. Phys. **18**, 75 (2022).
[48] S. T. Flammia and J. J. Wallman, Efficient estimation of Pauli channels, ACM Trans. Quantum Comput. **1**, 1 (2020).
[49] R. Harper, S. T. Flammia, and J. J. Wallman, Efficient learning of quantum noise, Nat. Phys. **16**, 1184 (2020).

[50] E. Robeva and A. Seigal, Duality of graphical models and tensor networks, Inf. Inference: A J. IMA **8**, 273 (2018).
[51] H.-Y. Huang, R. Kueng, and J. Preskill, Efficient estimation of Pauli observables by derandomization, Phys. Rev. Lett. **127**, 030503 (2021).
[52] T. Farrelly, N. Milicevic, R. J. Harris, N. A. McMahon, and T. M. Stace, Parallel decoding of multiple logical qubits in tensor-network codes, Phys. Rev. A **105**, 052446 (2022).
[53] S. Bravyi, M. Suchara, and A. Vargo, Efficient algorithms for maximum likelihood decoding in the surface code, Phys. Rev. A **90**, 032326 (2014).
[54] O. Kern, G. Alber, and D. L. Shepelyansky, Quantum error correction of coherent errors by randomization, Euro. Phys. J. D **32**, 153 (2005).
[55] E. Knill, Quantum computing with realistically noisy devices, Nature **434**, 39 (2005).
[56] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, Phys. Rev. A **94**, 052325 (2016).
[57] R. Harper and S. T. Flammia, Estimating the fidelity of T gates using standard interleaved randomized benchmarking, Quantum Sci. Technol. **2**, 015008 (2017).
[58] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B **7**, S347 (2005).
[59] F. Arute, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).
[60] K. C. Miao, *et al.*, Overcoming leakage in scalable quantum error correction (2022).
[61] M. McEwen, *et al.*, Removing leakage-induced correlated errors in superconducting quantum error correction, Nat. Commun. **12**, 1761 (2021).
[62] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, Scalable quantum circuit and control for a superconducting surface code, Phys. Rev. Appl. **8**, 034021 (2017).
[63] C. Chamberland, K. Noh, P. Arrangoiz-Arriola, E. T. Campbell, C. T. Hann, J. Iverson, H. Putterman, T. C. Bohdanowicz, S. T. Flammia, A. Keller, G. Refael, J. Preskill, L. Jiang, A. H. Safavi-Naeini, O. Painter, and F. G. Brandão, Building a fault-tolerant quantum computer using concatenated cat codes, PRX Quantum **3**, 010329 (2022).
[64] J. Napp and J. Preskill, Optimal Bacon-Shor Codes, Quantum Info. Comput. **13**, 490 (2013).
[65] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory **37**, 145 (1991).
[66] D. K. Tuckett, Ph.D. thesis, School of Physics, University of Sydney (2020), (QECSIM is available at https://github.com/qecsim/qecsim).
[67] E. v. d. Berg, Z. K. Minev, A. Kandala, and K. Temme, Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors (2022).
[68] R. Harper, I. Hincks, C. Ferrie, S. T. Flammia, and J. J. Wallman, Statistical analysis of randomized benchmarking, Phys. Rev. A **99**, 052350 (2019).
[69] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).
[70] J. Cotler and F. Wilczek, Quantum overlapping tomography, Phys. Rev. Lett. **124**, 100401 (2020).
[71] R. Harper, W. Yu, and S. T. Flammia, Fast estimation of sparse quantum noise, PRX Quantum **2**, 010322 (2021).
[72] S. T. Flammia and R. O'Donnell, Pauli error estimation via Population Recovery, Quantum **5**, 549 (2021).
[73] S. T. Flammia, Averaged circuit eigenvalue sampling (2021).
[74] T. Wagner, H. Kampermann, D. Bruß, and M. Kliesch, Pauli channels can be estimated from syndrome measurements in quantum error correction, Quantum **6**, 809 (2022).
[75] M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, S. T. Flammia, and A. S. Dzurak, Nonexponential fidelity decay in randomized benchmarking with low-frequency noise, Phys. Rev. A **92**, 022326 (2015).
[76] R. W. Andrews, C. Jones, M. D. Reed, A. M. Jones, S. D. Ha, M. P. Jura, J. Kerckhoff, M. Levendorf, S. Meenehan, S. T. Merkel, A. Smith, B. Sun, A. J. Weinstein, M. T. Rakher, T. D. Ladd, and M. G. Borselli, Quantifying error and leakage in an encoded Si/SiGe triple-dot qubit, Nat. Nanotechnol. **14**, 747 (2019).
[77] C. Gidney, Stim: a fast stabilizer circuit simulator, Quantum **5**, 497 (2021).
[78] M. A. Nielsen, A simple formula for the average gate fidelity of a quantum dynamical operation, Phys. Lett. A **303**, 249 (2002).
[79] J. M. Hammersley and P. Clifford, (1971), available at http://www.statslab.cam.ac.uk/ grg/books/hammfest/hamm-cliff.pdf.
[80] P. Abbeel, D. Koller, and A. Ng, Learning factor graphs in polynomial time & sample complexity, J. Mach. Learn. Res. **7**, 1743 (2006).