# Logical accreditation: a framework for efficient certification of fault-tolerant computations

James Mills[1],* Adithya Sireesh[1], Dominik Leichtle[1], Joschka Roffe[1], and Elham Kashefi[1,2]
[1]*School of Informatics, University of Edinburgh, United Kingdom*
[2]*Laboratoire d'Informatique de Paris 6, Sorbonne Université, France*

As fault-tolerant quantum computers scale, certifying the accuracy of computations performed with encoded logical qubits will soon become classically intractable. This creates a critical need for scalable, device-independent certification methods. In this work, we introduce logical accreditation, a framework for efficiently certifying quantum computations performed on logical qubits. Our protocol is robust against general noise models, far beyond those typically considered in performance analyses of quantum error-correcting codes. Through numerical simulations, we demonstrate that logical accreditation can scalably certify quantum advantage experiments and indicate the crossover point where encoded computations begin to outperform physical computations. Logical accreditation can also find application in evaluating whether logical circuit error rates are sufficiently low that error mitigation can be efficiently performed, extending the entropy benchmarking method to the regime of fault-tolerant computation, and upper-bounding the infidelity of the logical output state. Underpinning the framework is a novel randomised compiling scheme that converts arbitrary logical circuit noise into stochastic Pauli noise. This scheme includes a method for twirling non-transversal logical gates beyond the standard T-gate, resolving an open problem posed by Piveteau et al. [1]. By bridging fault-tolerant computation and computational certification, logical accreditation offers a practical tool to assess the quality of computations performed on quantum hardware using encoded logical qubits.

## I. INTRODUCTION

Experimental realisations of quantum error correction have reached a turning point [2–4]. With the ongoing improvement of physical qubit quality and system scale, we are beginning to enter the regime where target computations can be encoded into logical qubits. However, we are still far from realising the large-scale, fully fault-tolerant architectures required for universal computation using high-distance codes and complete fault-tolerant gate sets. For the foreseeable future, we will be operating in an intermediate regime of *early fault-tolerance*, where resource constraints limit both the error-correcting codes and the fault-tolerant protocols that can be implemented [5, 6]. In this regime, it can be unclear whether encoding computations in logical qubits actually improves performance over direct computation with physical qubits, given that magic state purification and logical noise suppression at scale are not yet feasible. This uncertainty presents a critical challenge for quantum error correction development: how can we assess whether fault-tolerant computations are actually trustworthy under realistic conditions?

This challenge is not exclusive to the near term. Even as large-scale fault-tolerant quantum devices become available, many assumptions underpinning quantum error correction, e.g. local, stochastic, or Markovian noise models, will remain problematic. For example, correlated errors were observed in recent surface code experiments in superconducting qubit hardware, revealing a logical error floor that prevents further error suppression [2]. These factors limit the ability of classical simulations and analytical models to certify computational correctness, motivating the need for alternative, device-independent methods of certification. How can we be confident that these assumptions hold in practice, or that our logical-level computations remain correct under deviations from particular noise models? What is needed is a new layer in the QEC stack: a tool for certifying that encoded quantum computations perform as intended, even in the presence of general, and possibly highly correlated, noise.

This paper introduces such a tool by bridging the fields of quantum error correction and computational certification. Specifically, we introduce *logical accreditation* as a scalable framework for efficient certification of computations performed with logical qubits encoded using stabiliser codes. Our protocol brings techniques from the field of certification (also known as verification or accreditation) [7], and extends them to the fault-tolerant regime by applying them to logical circuits implemented using stabiliser codes. The sampling overhead of logical accreditation is independent of both the number of logical qubits and the circuit depth. This inherent scalability means that the framework can be applied to certify computational correctness even in regimes that exceed the capabilities of the most advanced classical simulators.

A schematic overview of the protocol is shown in Fig. 1. Logical accreditation works by embedding the target computation within a larger ensemble of *trap* computations. The trap circuits are designed to be structurally identical to the target computation, but are configured to have deterministic outputs in the absence of noise. This deterministic behaviour of the trap circuits can be leveraged to provide a rigorous upper bound on the total variational distance (TVD) between the noisy and ideal

* J.Mills-7@sms.ed.ac.uk

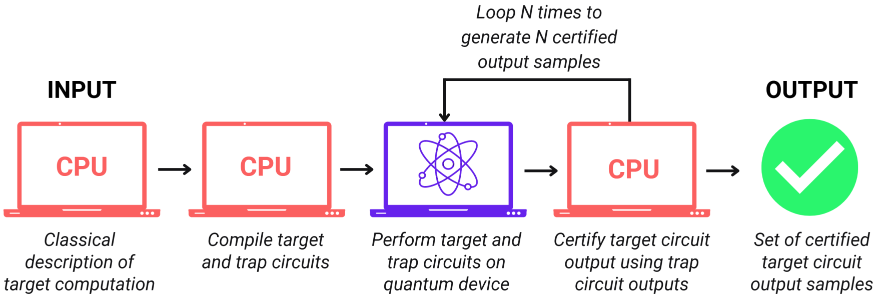

Loop N times to generate N certified output samples
INPUT
CPU
Classical description of target computation
CPU
Compile target and trap circuits
Perform target and trap circuits on quantum device
CPU
Certify target circuit output using trap circuit outputs
OUTPUT
Set of certified target circuit output samples

FIG. 1: The logical accreditation protocol is run jointly on a quantum device and a classical processing unit (CPU). The input is a classical description of the target computation. This description is used to compile both the target and trap circuits according to the required structure and gate set. These circuits are executed on the quantum hardware, and the resulting bit strings are recorded. The trap circuit outputs are then post-processed on the CPU to certify the corresponding target output. To obtain $N$ certified samples, the sampling procedure is repeated $N$ times. The final output is a certified set of $N$ bit string samples from the target circuit.

output distributions of the target computation.

A key advantage of logical accreditation is its ability to certify the accuracy of a *specific* computation performed on quantum hardware. This distinguishes it from methods that measure average performance, like the gate fidelity provided by randomised benchmarking [8–10] or the holistic device performance measured by quantum volume [11–13]. Crucially, logical accreditation provides these guarantees under far more general noise assumptions than are typically considered in the analysis of quantum error correction protocols.

One technical contribution of this work is a compilation scheme, inspired by randomised compiling [14, 15], that transforms general logical errors into tractable stochastic Pauli noise. Within this scheme, we introduce a novel method that resolves the open question of how to twirl non-transversal logical gates beyond the $T$ gate [1].

We demonstrate the power of the logical accreditation framework through numerical simulation of instantaneous quantum polynomial-time (IQP) and Trotterised circuits, also using noisy intermediate-scale quantum (NISQ) accreditation to certify physical unencoded computations [16]. This enables a direct comparison that indicates the crossover point where fault-tolerance delivers a practical advantage over unencoded quantum computation. The framework can also be used to extend entropy density benchmarking [17, 18] to the fault-tolerant regime, asses whether quantum error mitigation techniques can be efficiently applied to specific logical circuits, and provide an upper-bound on the infidelity of the output state of a logical computation. This work builds on ideas from the family of protocols known as accreditation, originally developed to certify circuits executed on noisy physical qubits in both the digital and analog settings [16, 19–21], as well as from cryptographic verification protocols [7, 22, 23].

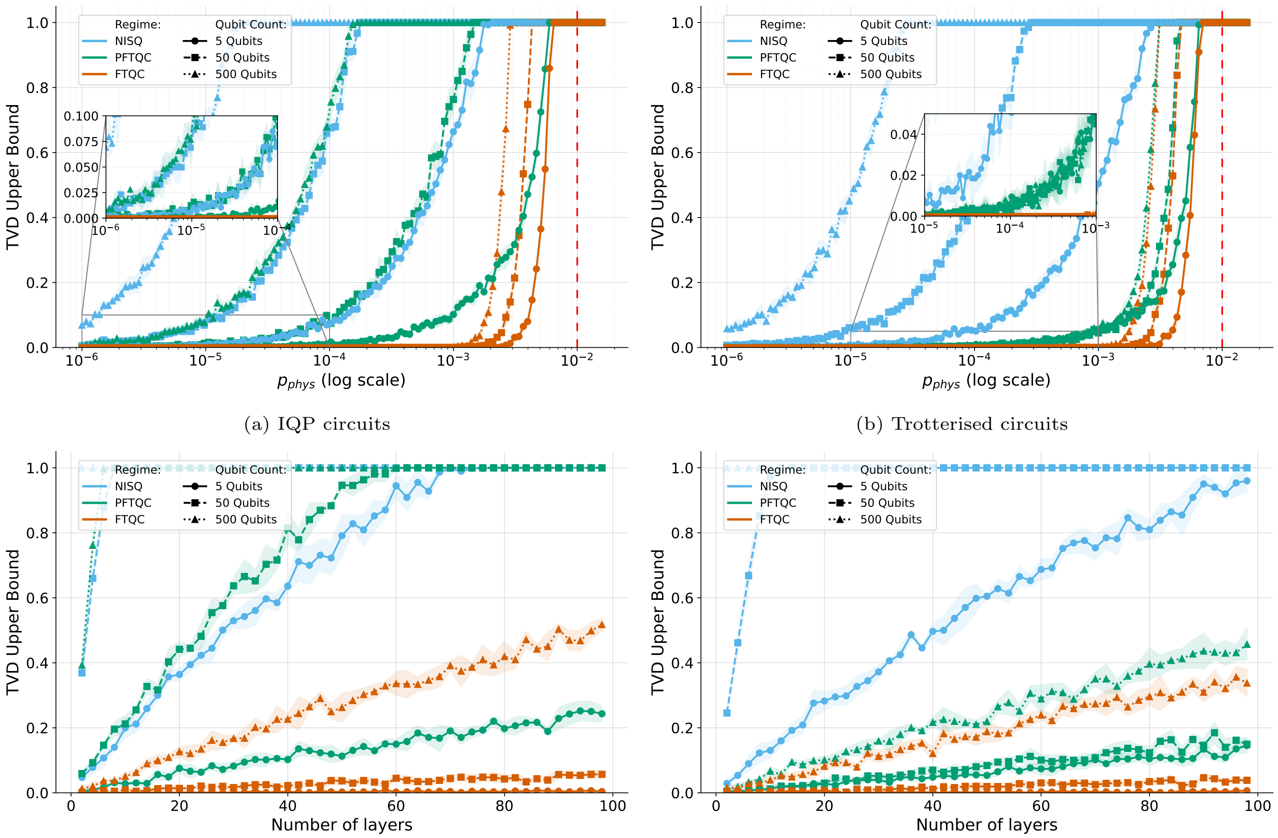
## II. PRELIMINARIES

### A. Computation with logical qubits encoded using stabiliser codes

An $[[n, k, d]]$ stabiliser code encodes $k$ logical qubits using $n$ physical qubits, with a code distance $d$, where the code distance is the minimum weight with which any non-trivial logical operator acts. A stabiliser group $\mathcal{S}$ is a commutative subgroup of the Pauli group not containing the operator $-I^{\otimes n}$, or, in other words, its elements form an Abelian subgroup of the Pauli group. The stabiliser group has a minimal representation in terms of a set of independent generator elements of size $n-k$ i.e. $\mathcal{G} = \{g_1, \dots, g_{n-k}\}$. The simultaneous $+1$ eigenspace of a given stabiliser group defines the codespace of a QEC code. The codespace is then a $2^k$-dimensional subspace of the larger $2^n$-dimensional Hilbert space, while the remaining $2^{n-k}$-dimensional subspace is the error-space. A logical basis of the form $\{|0\rangle_L, |1\rangle_L\}^{\otimes k}$ can be defined in the codespace, along with logical Pauli operators that act on the encoded qubits.

If a state, $|\psi\rangle_L$, is within the codespace of the code, then $P_i |\psi\rangle_L = |\psi\rangle_L$ if $P_i \in \mathcal{S}$. During cycles of error correction, $n-k$ stabiliser measurements are made, using the elements of the generator group $\mathcal{G}$. Depending on whether errors have occurred, each stabiliser measurement projects the logical state onto either the $+1$ or $-1$ eigenspace of the measurement operator, where the measurement projector is $P_{g_i,s} = \frac{1}{2}(I + (-1)^s g_i)$ for $g_i \in \mathcal{G}$ and syndrome measurement output $s \in \{0, 1\}$. If the syndrome measurement outputs of all the stabiliser generators are 0, this indicates the state is in the codespace. If any syndrome measurements output 1 this indicates the state is in the error-space. Repeated rounds of syndrome measurements, with corrections either applied or tracked as errors occur, can be used to protect quantum information stored in the logical codespace for use as a quantum memory.

To perform computation with logical states, it is neces-



(a) IQP circuits

(b) Trotterised circuits

(c) IQP circuits

(d) Trotterised circuits

FIG. 2: *Numerical simulations applying logical accreditation to IQP and Trotterised circuits.* Circuits are run on noisy physical qubits (NISQ), on logical qubits without magic state purification (PFTQC), and with purification (FTQC). Logical error rates are simulated using surface code parameters with code distance 11. NISQ circuit certification uses the method from [16], while logical circuits use logical accreditation. In (a) and (b), the certified TVD upper bound is plotted against physical error rate for IQP and Trotterised circuits respectively. In (c) and (d), TVD is plotted against the number of circuit layers for each case. The physical error rate is fixed at $p_{\text{phys}} = 10^{-3}$, with 500 traps used per protocol run.

sary not only to reliably store quantum information in the codespace but also to apply logical gate operations. For a given quantum error-correcting code, a subset of logical gates known as transversal gates can be implemented fault-tolerantly without incurring additional qubit overhead. The Eastin-Knill theorem states that no QEC code has a set of transversal gates with which universal computation can be performed [24].

In the logical accreditation framework, we assume that qubits are encoded with a stabiliser code where Clifford operations may be performed transversally, or using alternative fault-tolerant methods like lattice surgery [25]. Non-Clifford gates, required for universal computation, are performed through the consumption of magic states using gate teleportation or projective measurement. Magic states are quantum states prepared using ancillary logical qubits, and in this work, will be assumed to be of the form

$$|\theta\rangle := \frac{1}{\sqrt{2}}\big(\,|0\rangle + e^{i\theta}\,|1\rangle\,\big). \tag{1}$$

Gate teleportation or projective measurement operations involving magic states can be used to perform Pauli rotation operations on the computational logical qubits.

If a magic state is phase rotated by angle $\theta = m \cdot \pi/4$ for $m \in \mathbb{Z}$, then a single round of gate teleportation may be used to apply the desired gate. While if $\theta \neq m \cdot \pi/4$, a repeat-until-success (RUS) approach may instead be used [26, 27]. In RUS gates, the correct rotation by angle $\theta$ is applied with probability 1/2. If the incorrect rotation is applied, then a new magic state is prepared instead rotated by $2\theta$, and the gate is applied again; this process is repeated until the intended rotation is achieved. The average number of repetitions for RUS gate success is

$$1 \times \frac{1}{2} + 2 \times \frac{1}{4} + \ldots = \sum_{i=1}^{\infty} \frac{i}{2^i} = 2. \tag{2}$$

A projective Pauli measurement approach can be applied to use magic states to perform high-weight logical Pauli rotations. Projective measurements may be performed fault-tolerantly using lattice surgery, and the number of high-weight logical rotation operations can be optimised during circuit compilation [28].

### B. Fault-tolerant quantum computation

A logical qubit encoded using an $[[n, k, d]]$ stabiliser code can tolerate up to $\lfloor (d-1)/2 \rfloor$ physical errors. If this bound is exceeded, the decoder may apply an incorrect correction, resulting in a logical error. Logical errors are unwanted operations that alter the encoded state without taking it out of the code space. Because they do not produce a detectable syndrome, they are not corrected during rounds of error correction.

The space-time constraints of early fault-tolerant quantum devices will necessitate trade-offs that balance error suppression with computation depth. These constraints will mean that logical error rates cannot be arbitrarily reduced as the scale of computations increases. Factors contributing to the overhead of logical computation are the scale of the computation being performed, the code distance of the logical qubits, and the extent to which magic states are purified. Decreasing the code distance and amount of purification will allow larger computations to be performed, but at the cost of increasing logical error rates. This is because lower code distances decrease the number of qubits that need to be affected by errors before a logical error can occur, while reducing the amount of purification used to improve the quality of the magic states leads to noise being introduced into the logical computation every time a non-transversal non-Clifford gate is performed.

Computational frameworks suitable for early fault-tolerant devices have been proposed in which magic states are not purified, but are prepared and immediately used to perform noisy logical non-Clifford operations [1, 26].

We will distinguish between two regimes of logical computation: *partial fault-tolerance* and *full fault-tolerance*. Following the convention used in [29] and [26], we will use the term partial fault-tolerance to mean logical computation performed without purification of magic states, while by full fault-tolerance, we refer to logical computation where magic state purification is implemented so that logical non-Clifford gates are performed with approximately the same error rates as logical Clifford gates. In Appendix I, we describe the Clifford plus $T$ framework for fault-tolerant computation and the Clifford plus noisy $T$ framework for partially fault-tolerant computation. We also describe an alternative method for partially fault-tolerant computation that uses a gate set composed of Cliffords and noisy analog rotations.

## III. THE LOGICAL ACCREDITATION PROTOCOL

### A. Overview

Logical accreditation is a framework for certifying fault-tolerant computations, where logical qubits are encoded using stabiliser QEC codes. A simplified schematic for the protocol is shown in Fig. 1.

The input to the logical accreditation protocol is a classical description of the computation to be run on the quantum device. In the first step of the protocol, the input computation is compiled into a logical circuit, termed the *target computation.* The target circuit is compiled using a repeating structure of gate layers containing single- or multi-qubit non-Clifford gates and CZ gates, sandwiched between gate layers containing single-qubit Clifford gates. Then $M$ distinct logical *trap circuits* are compiled. These circuits are designed to mirror the structure of the target circuit but produce a known deterministic bit-string output in the absence of noise.

The logical target and trap circuit structures are shown in Fig. 3 (a) and (b), respectively.

The construction of the trap computations is described in detail in Appendix A.

Within the framework, a circuit compilation scheme termed logical randomised compiling is introduced to twirl all logical circuit noise into logical stochastic Pauli noise. If one of the trap circuits is affected by noise, their construction guarantees that the ideal output is not measured with a probability lower bounded by a constant, as described in Appendix D and E.

During logical accreditation, the target and trap circuits are run in a random order on the quantum device using logical qubits encoded using a stabiliser code with the chosen logical gate set. The measurement outcomes of the traps are then used to certify the quality of the target circuit. Specifically, the accuracy of the target circuit output is quantified in terms of a bound on the total variation distance (TVD) between the experimental target circuit output distribution, $\mathcal{D}_{exp}$, and its ideal output distribution, $\mathcal{D}_{ideal}$.

### B. Noise assumptions

The logical accreditation protocol makes the following assumptions on the structure of the noise affecting the quantum device:

(A1) Physical single-qubit gate noise is gate-independent, but can depend on the gate position.

(A2) Logical noise affecting single-qubit logical Clifford gates, due to uncorrected noise during error correction cycles, is gate-independent but can depend on the gate position.

(A3) Noise affecting logical qubits, due to uncorrected noise during error correction cycles, is modelled by completely positive trace-preserving (CPTP) maps acting at the logical level, and noise affecting distinct logical circuit runs is independent.

The above assumptions are mild in the sense that they allow for very general noise to affect the logical circuits during the protocol. This can be noise that is non-local, and highly correlated, including the most common types of quantum device noise such as cross-talk effects, qubit decoherence and amplitude damping, and gate calibration errors.

The robustness of the method to violation of these assumptions is discussed in Section 7.

Assumption A1 may be physically justified as single-qubit gates generally have the lowest error rates compared to other quantum operations, making variation between their noise profiles marginal. If single-qubit logical Clifford gates are performed transversally using only physical single-qubit gates, then assumption A2 follows as a consequence of assumption A1. The independence assumption of A3 is used to guarantee the convergence of the bound provided by logical accreditation. This might be relaxed if some other type of convergence guarantee is provided. Whether the noise is assumed to be Markovian or non-Markovian determines which soundness bound is used for the protocol; this is discussed in Section III E.

Approximations made during compilation can be another source of error. Approximation errors can occur if, for example, the Solovay-Kitaev algorithm is used to approximate arbitrary unitary gates to within an operator norm distance of $\epsilon > 0$ from the ideal unitary, with a gate sequence of length $O(\log(1/\epsilon))$ using only a discrete gate set [30, 31]. Compilation errors will also occur when an approximation methods such as Trotterization are used to simulate the evolution of a Hamiltonian with time[32–34]. While the logical accreditation framework is compatible with these methods, it does not detect errors caused by such approximations. However, bounds on approximation errors of the kinds mentioned may often be efficiently computed using classical methods, allowing them to be straightforwardly accounted for.

### C. Certification protocol

We now describe the main result of this work. In the logical accreditation protocol, a target circuit and $M$ trap circuits are performed with a random ordering on a quantum device using encoded logical qubits. During the protocol, logical noise twirling is applied to the target and trap circuits using a random compilation scheme as described in Appendix B 1. This compilation is designed to convert general logical circuit noise from incorrect decoding or from imperfect magic state preparation into logical stochastic Pauli noise. This means the effective state generated by running the target circuit or a trap circuit may be decomposed as the convex combination

$$\rho_{out}^{(i)} = (1 - p_{err})\rho_{out,id}^{(i)} + p_{err}^{(i)}\rho_{noisy}^{(i)}, \tag{3}$$

where the index $i \in \{1, ..., M+1\}$ indicates which circuit has been run to generate the quantum state, $p_{err}^{(i)}$ is the probability of at least one logical error occurring at any point in the circuit, $\rho_{out,id}^{(i)}$ is the ideal output state in the absence of noise, and $\rho_{noisy}^{(i)}$ is the output state encompassing the effects of logical circuit noise.

The trap circuits are designed to deterministically output a known bit string in the absence of noise. If a trap circuit is run on the quantum device and the measured output bit string differs from the ideal output, this is recorded as a trap circuit failure. The logical noise information provided by the trap circuits is used to upper-bound the total variation distance (TVD) of the experimental output of the target circuit from the ideal output.

The framework requires a classical description of the target computation as input. In the first step of the protocol, the target computation is compiled into a logical circuit using a gate set that can include the following logical gate operations: single-qubit Clifford gates, $CZ$ gates, and Pauli-operator rotation gates of arbitrary weight and rotation angle. The non-Clifford gates are performed by consuming magic states using a gate teleportation or projective measurement approach. A repeating circuit structure is used for the circuit compilation, consisting of layers of single-qubit Clifford gates and layers of multi-qubit Clifford gates and non-Clifford single- or multi-qubit Pauli rotation gates. Each logical qubit within one logical gate layer is acted on by at most one logical gate. Next, $M$ randomly constructed trap circuits are generated. These circuits replicate the structure of the target circuit, but with all non-Clifford components replaced by Clifford operations. In particular, each state injection routine — such as those used to implement $T$-gates — is modified to inject a logical identity state instead. This substitution preserves the error propagation behaviour of the target circuit while ensuring that each trap circuit yields a deterministic output. The target and trap circuit structures are shown in Fig. 3, and the construction of trap circuits is described in more detail in Section III F.

The target and trap circuits are run on the quantum device with a random ordering, and their measured bit string outputs are recorded. With their deterministic outputs, the trap circuits provide the reference for computing the output of the logical accreditation protocol: an upper bound on the TVD between the experimental and ideal output distributions.

**Theorem 1.** *The logical accreditation protocol provides an upper-bound on the total variational distance (TVD) between the experimental target circuit output distribution and the ideal output distribution of the form*

$$\frac{1}{2}\sum_{s\in\{0,1\}^n} |p_{exp}(s) - p_{ideal}(s)| \leq \gamma, \tag{4}$$

*where $\gamma$ is experimentally estimated from the measured outputs of the logical trap computations. The bound applies with a user defined accuracy $\epsilon$ and confidence $\alpha$, where the number of trap circuits satisfies the inequality $M > \frac{2}{\epsilon^2}\log\left(\frac{4}{1-\alpha}\right)$. The assumptions necessary for the protocol are the previously stated conditions* A1-A3.

The derivation of this bound can be found in Appendix C. Logical accreditation can be repeated an arbitrary number of times to generate a certified set of output bit strings for the target computation, as shown in Fig. 1. In this way, the protocol can be applied to compute distributions and expectation values with certified error bounds. The error of the expectation value of an observable $O$ is bounded by

$$|\langle O\rangle_{\rho_{exp}} - \langle O\rangle_{\rho_{id}}| \leq 2\gamma\|O\|, \tag{5}$$

where $\|.\|$ is the operator norm, also called the spectral norm, $\langle O\rangle_{\rho_{exp}}$ is the experimentally estimated expectation value, and $\langle O\rangle_{\rho_{id}}$ is the ideal expectation value. The logical accreditation bound is also an upper bound for the infidelity, as:

$$1 - F(\rho_{out,id}, \rho_{out}) \leq \gamma, \tag{6}$$

where $F(.,.)$ denotes the fidelity between two quantum states, and $\rho_{out,id}$ and $\rho_{out}$ are the ideal and experimentally produced target circuit output states; this result is derived in Appendix L, where there is also numerical analysis comparing the eqn. 6 bound with the infidelity of states generated by IQP circuits is plotted in Fig. 18.

### D. Robustness

If the previously stated noise assumptions are violated, the bound computed by the protocol is then of the form $\gamma' = \gamma + \epsilon_d$, where $\gamma$ is the upper bound provided by the protocol if the assumptions hold, $\gamma'$ is the upper-bound provided by the protocol if the assumptions are violated, and $\epsilon_d$ is the difference in the protocol bound induced by violation of the assumptions.

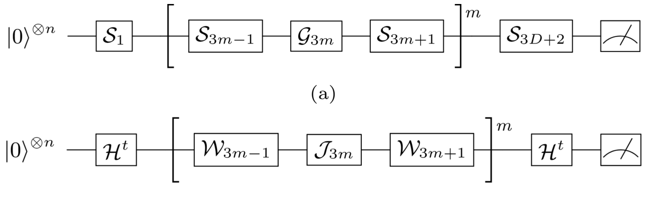
The following result is derived in Appendix G:

**Theorem 2.** *If the protocol assumptions are violated, the absolute difference in the computed upper-bound, $\gamma'$, from the upper-bound computed where the assumptions hold, $\gamma$, is bounded*

$$|\gamma - \gamma'| \leq M^{-1}\sum_k\sum_j ||\mathcal{E}_j^{(k)} - \mathcal{E}_j^{(k)}{'}||_\diamond, \tag{7}$$

*where $\{\mathcal{E}_j^{(k)}\}_j$ and $\{\mathcal{E}_j^{(k)}{'}\}_j$ are set of noise channels affecting the $k$-th trap circuit where the protocol assumption hold, and the set of noise channels affecting the $k$-th trap circuit where these assumptions are violated, respectively.*

This implies that any deviation in the bound due to violation of the assumptions depends only linearly on the diamond norm distance of each of the corresponding noise channels affecting each circuit. And if the assumptions are only weakly violated, i.e. each term $||\mathcal{E}_j^{(k)} - \mathcal{E}_j^{(k)}{'}||_\diamond$ is small for all $j$ and $k$, the outputs of the trap circuits remain close to their outputs where the assumptions hold. Therefore, the protocol is robust to any small violations of the assumptions, since these will have a small effect on the computed upper-bound.



(a)

(b)

FIG. 3: *The circuit structure used for compilation of target and trap circuits, including state preparation and measurement.* The logical gate layers are numbered by subscript with their ordering in time. The bracketed parts of the circuits indicate a repeated structure, with $m \in \{1,\dots,D\}$, where the gates within each of the layers can change between repetitions. The circuit diagram in (a) shows the structure used to compile the target circuit, where the notation $\mathcal{S}$ denotes a layer of single-qubit Clifford gates, $\mathcal{G}$ a gate layer that can contain multi-qubit Clifford gates, non-Clifford gates and identity gates. The circuit diagram in (b) shows the trap circuit structure used for compilation, where $\mathcal{H}$ denotes a layer of Hadamard gates, with $t \in \{0,1\}$, $\mathcal{W}$ a single-qubit randomising gate layer containing $S$, $S^\dagger$ and $H$ gates, and $\mathcal{J}$ a gate layer that can contain multi-qubit Clifford gates, trap circuit versions of the corresponding non-Clifford gates found in the target circuit, and identity gates. The logical randomised compiling gate layers are omitted from these diagrams for the sake of simplicity.

### E. Completeness and soundness

If the measured output of a trap circuit deviates from the ideal output, a logical error is guaranteed to have occurred. Therefore, the false negative rate of the protocol is 0, and it has perfect completeness. There are two factors to be considered for the soundness of the protocol: (1) The probability that if an error affects a trap circuit that it will be detected. (2) The probability that if multiple errors affect a trap circuit, they cancel and so are not detected. In Appendix D, we show that any single logical Pauli error of arbitrary weight occurring at any single time-step during a trap circuit is detected with probability $p_{det} \geq 1/2$. We derive upper-bounds on the probability of error cancellation for three different scenarios. Non-Markovian noise can result in classically correlated errors that may cancel in the trap circuits. The three scenarios reflect different assumptions that can be made about the Markovianity of the noise. For each scenario, an upper-bound on the probability of error cancellation is provided. First, we show that if logical error rates are sufficiently low and the noise is Markovian then error cancellation can be neglected. Second, if the dominant sources of error are multi-qubit Clifford gates and gates requiring magic states, then the probability of error cancellation is bounded by $p_{canc} \leq 1/2$. This bound is suitable under the assumption that the dominant gate noise is non-Markovian, i.e. the noise from multi-qubit Clifford gates and gates requiring magic states, while noise from single-qubit Clifford gates is Markovian. And third, we show that with a slight modification to the trap circuit construction, the probability of any error cancellation is bounded by $p_{canc} \leq 7/8$. For this result, no assumptions are made about the Markovianity of the noise. These

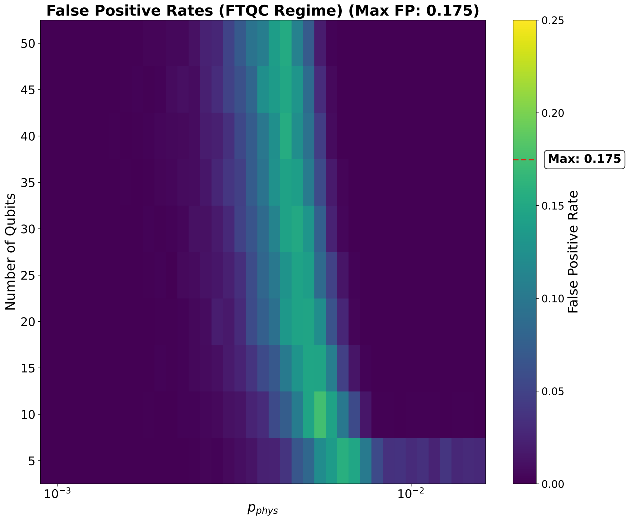


FIG. 4: *Soundness heatmap for IQP circuit sampling.* This shows the false positive rates for different numbers of qubits and error rates when performing IQP sampling in the full fault-tolerance (FTQC) regime. A logical depolarising noise model was used for the numerical simulations, with the false positives occurring due to stabilisation and error cancellation. The observed false positive rates are well below the analytical soundness bounds given in Section III E

results are derived in Appendix E.

Combining the bound for error detection probability with each of the three bounds for error cancellation probability yields three different soundness parameters: $\delta_1 = 1/2$, $\delta_2 = 3/4$, and $\delta_3 = 15/16$. These soundness parameters upper bound the false positive rate of the trap circuits, and may be used to derive different TVD bounds. This is described in Appendix C. Fig. 4 shows a heatmap of false positive rates for different numbers of qubits and physical error rates when performing IQP circuit sampling in the regime of full fault-tolerance. The simulations were run for 40-layer circuits, varying the number of logical qubits from 5 to 50 and the physical error rate from $10^{-3}$ to $5 \times 10^{-2}$. In all cases, the experimentally observed false positive rates are well below the theoretical soundness bounds.

### F. Trap computation construction

The randomly generated trap circuits have the same structure as the target circuit, and stabilise the initialised logical state in the absence of logical errors. Each of the logical gate operations performed during a trap circuit either compiles to an identity operation, or to an operation that stabilises the input logical state. This means the trap circuits provide a deterministic bit string output in the absence of noise.

The positioning and type of the multi-qubit Clifford gates are kept the same in the trap circuits as they are for the target circuit, but are sandwiched uniformly at random by $S$ and $H$ logical gates. The overall logic of the randomly applied $S$ and $H$ operations, combined with the multi-qubit Clifford gates, results in randomly oriented CNOT gates. A layer of logical Hadamard gates is included at the beginning and end of the trap circuits with probability 1/2, otherwise layers of identity operations are applied. The randomly chosen $S$ and $H$ gate layers prevent the same Pauli noise from cancelling in different trap circuits, meaning the trap circuits can detect logical Pauli errors affecting the trap circuits with constant lower-bounded probability.

In the trap circuits, modified versions of the corresponding non-Clifford gates in the target circuit are performed using $\pi$-rotated magic states. These are then immediately followed by fault-tolerant Pauli gates, with the result that, in the absence of noise, the overall logic of each of these operations is identity. If no magic state purification is used during the target and trap computations, the protocol's overhead consists solely of an additional sampling cost, which scales independently of the system size. Two versions of each trap circuit are run on the quantum device. The first uses only magic states of the form $|\pi\rangle$, and the second only of the form $|\pi/2\rangle$. In the second case, a fault-tolerant $S$ gate is applied immediately after the magic state has been prepared so that the final trap magic state is a $|\pi\rangle$ state. If either version of the trap circuit records an error when run on the quantum device, this is recorded as failure for that trap circuit. Together, the $|\pi\rangle$ and $|\pi/2\rangle$ magic states are vulnerable to all possible logical Pauli errors that might affect the magic states during a potentially imperfect state-preparation procedure. If purification methods are used to improve the quality of the magic states, a different procedure is used to generate the trap magic states. Without requiring additional assumptions, magic state purification is included in the protocol at the cost of a factor of 2 increase in the number of magic states consumed. Gate teleportation is performed on pairs of purified $|\pi/4\rangle$ magic states to generate $|\pi/2\rangle$ states, which are then transformed into $|\pi\rangle$ states by applying fault-tolerant $S$ gates. However, if it is assumed that $|\pi/4\rangle$ and $|\pi\rangle$ state purification generates states of the same quality, then no additional sampling or physical resources are required. If some of the non-Clifford gates in the target circuit are performed using purified magic states and some not, then this requires both the additional sampling overhead for the non-purified magic states and the physical overhead accompanying the purified states.

Under the previously stated noise assumptions, this trap circuit construction ensures that both the target and trap circuits are subject to the same distribution of possible noise channels. More details are provided in Appendix A.

### G. Logical randomised compiling

The logical accreditation protocol uses a compilation scheme used for logical noise twirling termed *logical ran-*

*domised compiling.* This may be seen as an extension of the NISQ circuit method of randomised compiling [14] to apply to encoded logical qubits and gates performed by the consumption of magic states. Logical twirling operations are directly integrated into the target and trap circuits without changing their computational logic. The logical Pauli gate twirling operations create effective noise channels, transforming general logical noise, including, for example, coherent logical noise that can affect logical qubits [35], into stochastic logical Pauli noise. This means that the logical circuits generate measured outputs as if they are subject to the effective stochastic Pauli channels generated by the twirling.

In this paper, we introduce a novel method for twirling logical arbitrary non-Clifford gates, addressing an open problem originally outlined in Pivateau et al. [1].

The method described in [1] for twirling $|\pi/4\rangle$ magic states using the gate set $\{I, e^{-i\pi/4}SX\}$ relied on the $T$ gate being in the third level of the Clifford hierarchy. The new method we propose requires only assumption A1, namely that physical single-qubit gate noise is gate-independent. It involves updating the magic state preparation procedure depending on the twirling operations applied. Our twirling method can be used to twirl logical Pauli rotation gates of arbitrary weight and rotation angle. A similar twirling approach can also be used to twirl arbitrarily rotated magic states, beyond the standard $\pi/4$ phase-rotated states. Details on the implementation of logical randomised compiling are provided in Appendix B.

There is evidence that for sufficiently high code distances, logical noise affecting qubits encoded using the surface code convergences towards logical stochastic Pauli noise [35]. However, if the number of available qubits is limited this will restrict the achievable code distance. This necessitates the use of logical randomised compiling. Recent work showed that it is possible to twirl over a subset of the Clifford group that commutes with a non-Clifford logical gate operation [36]. We note that the approach that we propose could be adapted to perform logical non-Clifford gate twirling over the full Clifford group.

## IV. NUMERICAL EXPERIMENTS

The ability to quickly assess the accuracy of applications executed on fault-tolerant quantum devices is essential to their practical utility. Instantaneous Quantum Polynomial (IQP) circuit sampling and Hamiltonian simulation are two tasks for which certification of computational accuracy is crucial when it comes to assessing whether a quantum advantage has been achieved [37–39]. In this paper, we apply the logical accreditation protocol to certify the accuracy of logical computations in numerical simulations of logical qubits encoded using the surface code for IQP circuit sampling and Hamiltonian simulation. This is described in the two subsections below.

### A. Simulation details

We numerically simulate partially and fully fault-tolerant circuits to test the logical accreditation framework. For comparison, we also use an existing NISQ accreditation protocol to certify computation on physical qubits [16].

For the simulations of partially and fully fault-tolerant circuits, we assume our computation is protected using surface code quantum error correction. Following [2], the logical error rate $p_L$ for the surface code is modelled by the expression

$$p_L \propto \left(\frac{p_{phys}}{p_{th}}\right)^{\frac{d+1}{2}} \tag{8}$$

where the physical error rate is denoted by $p_{phys}$, the threshold error rate is denoted by $p_{th}$, and the code distance is denoted by $d$. In all simulations, we assume the surface code threshold is $p_{th} = 0.01$, and use a constant prefactor of $c = 0.03$ [40].

Circuit noise is modelled using depolarising channels, applied either at the physical or logical level depending on the circuit type. For NISQ simulations, all gate operations are affected by depolarising noise with parameter $p_{phys}$. In partially fault-tolerant simulations, Clifford gates are assumed to be encoded and are subject to logical noise at rate $p_L$, while non-Clifford gates are assumed to rely on unpurified magic states and so are subject to physical noise at rate $p_{phys}$. In fully fault-tolerant simulations, all gates are subject to logical noise at rate $p_L$, regardless of type.

Each target computation is certified under all three frameworks:

1. NISQ accreditation using $n$ physical qubit circuits,
2. Logical accreditation with $n$ partially fault-tolerant logical qubit circuits,
3. Logical accreditation with $n$ fully fault-tolerant logical qubit circuits.

This yields three TVD bounds for direct comparison for each computation framework. By examining these bounds, we identify the operating regimes in which NISQ, partially fault-tolerant, or fully fault-tolerant computation is most effective. Unless otherwise stated, we use $M = 500$ trap circuits to compute the TVD bound. Each data point in our plots represents the mean and standard deviation of this bound, calculated over five independent runs, where 500 traps are independently sampled for each run.

### B. Certifying IQP circuit sampling

IQP circuit sampling is a restricted model of quantum computation believed to be hard to simulate classically

[37]. An $n$-qubit IQP circuit starts with all qubits in the $|0\rangle^{\otimes n}$ state, followed by a layer of Hadamard gates, a unitary made from randomly selected diagonal gates, another layer of Hadamard gates, and measurement in the computational basis. Approximating the output distribution, even in the presence of noise [38, 41], is #P-hard. This makes IQP circuits strong candidates for demonstrating quantum advantage on early fault-tolerant devices [39].

Logical accreditation can be used to certify that logical error rates are below the threshold required for quantum advantage. According to [41], if an IQP output distribution can be classically sampled to within 1-norm error of 1/192, then the Polynomial Hierarchy would collapse to the third level. Certifying a TVD that satisfies this threshold would therefore provide strong evidence of quantum advantage.

Fig. 5 shows how this threshold constrains the number of possible $T$ gates that can be used in a circuit such that the threshold is satisfied, for a range of $T$-gate noise rates, assuming perfect Clifford gates. The highlighted region indicates where quantum advantage may be achievable: error rates are sufficiently low and qubit counts are sufficiently high to make classical simulation infeasible.

Fig. 2 presents results from our simulations. In panel (a), we show certified TVD bounds for circuits with 40 gate layers and 5, 50, or 500 qubits. Each circuit is implemented in three versions: NISQ, partially fault-tolerant, and fully fault-tolerant. For the fault-tolerant circuits, we use distance-11 surface codes to model logical noise. This distance is chosen to explore a realistic noisy regime where the benefits of fault tolerance become clear. We find that as physical qubit error rates improve, the TVD bound decreases. Once the error rate drops below the code threshold, fault-tolerant circuits outperform unencoded ones. Fully fault-tolerant circuits perform best overall, followed by partially fault-tolerant circuits, and then NISQ.

In panel (c), we fix the physical error rate at $p_{phys} = 10^{-3}$ and vary the number of gate layers, keeping qubit counts and trap count fixed as before. As depth increases, TVD increases. NISQ circuits quickly approach TVD $\approx 1$, while partially and fully fault-tolerant circuits degrade more slowly. TVD also increases faster for circuits with more qubits.

We also explore how performance changes with code distance and magic state quality. Fig. 6 shows that increasing code distance (from 3 to 13) reduces logical error rates and tightens the TVD bound, but at the cost of higher resource overhead. Finally, we test how the fidelity of magic states affects performance. In Fig. 8 (a), we fix the physical error rate (and therefore the Clifford error rates) and vary only the $T$-gate error rate arising from magic state imperfections. We find that improving magic state fidelity helps reduce TVD, but only up to a point. For low code distances, the benefit saturates as Clifford noise becomes dominant.

In Appendix M, we estimate the minimum physical resources needed to achieve a certified TVD below the quantum advantage threshold for IQP sampling on superconducting devices using surface codes.

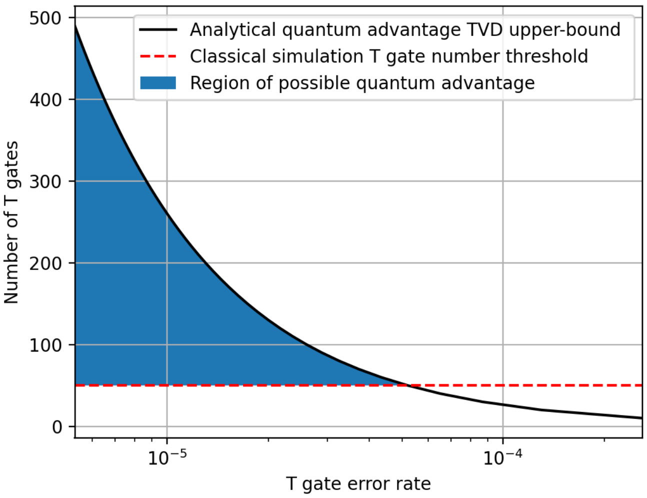


FIG. 5: *Region of possible quantum advantage for IQP circuit sampling depending on $T$ gate noise and number of $T$ gates.* The solid line on the plot represents the maximum number of noisy $T$ gates such that the quantum advantage threshold derived in [41] is satisfied, plotted as a function of the $T$ gate error rate. It is assumed that Clifford gates are error-free, and circuit noise originates solely from $T$ gates. The best-known classical algorithms for simulating quantum circuits can simulate quantum circuits with up to around 50 $T$ gates [42, 43]. This threshold number of $T$ gates is included as the dashed horizontal line on the plot. The shaded region between the classical simulation threshold and the noise threshold indicates experimental parameters where a quantum advantage is possible.

### C. Certifying Trotterised quantum Hamiltonian simulation

We now consider the problem of certifying quantum Hamiltonian simulation with Trotterised circuits. Hamiltonian simulation is considered a strong candidate application for fault-tolerant quantum computation. One Hamiltonian that would be interesting to simulate is the one-dimensional nearest-neighbour Heisenberg model with a random magnetic field in the $Z$-direction. This model is of broad interest, especially in the context of phenomena such as self-thermalisation and many-body localisation [44]. Due to its interest in the condensed matter community, it has been used to benchmark various methods of quantum simulation [45]. Hamiltonian evolution, such as that of the one-dimensional nearest-neighbour Heisenberg model, can be simulated on a quantum computer using techniques such as Trotterised time evolution [29]. Trotterisation has been found to display better relative performance, when applied to achieve a finite target precision, than other asymptotically better scaling techniques [46]. This makes it a promising method for simulating quantum evolution on early fault-tolerant devices [27, 29].

The Trotter-Suzuki decomposition allows the time evolution operator $e^{-iHt}$ to be approximately decomposed into a sequence of $N$ discrete time-steps [29, 46]. The second-order Trotter-Suzuki decomposition approximates these time-step evolution terms as

$$\left(e^{-i(\sum_{i=1}^{L} a_i P_i)t/N}\right)^N \approx \left(\prod_{i=1}^{L} e^{-i(\frac{a_i t}{2N})P_i} \prod_{i=L}^{1} e^{-i(\frac{a_i t}{2N})P_i}\right)^N . \tag{9}$$

If the Hamiltonian is expressible as a linear combination of Pauli operators, as is possible for the one-dimensional nearest-neighbour Heisenberg Hamiltonian, the previous expression may be written as a sequence of multi-qubit Pauli rotations

$$\begin{aligned}\left(\prod_{i=1}^{L} e^{-i(\frac{a_i t}{2N})P_i} \prod_{i=L}^{1} e^{-i(\frac{a_i t}{2N})P_i}\right)^N \\ = \left(\prod_{i=1}^{L} R_{P_i}\left(\frac{a_i t}{2N}\right) \prod_{i=L}^{1} R_{P_i}\left(\frac{a_i t}{2N}\right)\right)^N .\end{aligned} \tag{10}$$

The Trotterised circuit consists of a sequence of high-weight Pauli rotations defined by the above expression. To run this circuit using surface code logical qubits, each high-weight logical Pauli rotation gate may be performed via lattice surgery techniques, with each non-Clifford Pauli rotation requiring the consumption of a single magic state. Since the approximation error of Trotterisation may be efficiently bounded, the main source of error stems from logical noise when the algorithm is run on hardware.

Logical accreditation is compatible with Trotterised target circuits composed of logical multi-qubit Pauli rotation gate operations.

It can be used to derive an upper bound on the total variation distance (TVD) of the output when measuring an observable of the time-evolved state, as well as an error bound for the corresponding expectation values. Although most applications involving circuit Trotterisation have some computational steps after the Trotterised evolution, for example, the QCELS algorithm [29], for the sake of generality we only consider bounding the error of the Trotterised circuit. The approach used in the numerical experiments may straightforwardly be adapted for a particular application of Trotterisation.

We applied the logical accreditation framework in numerical simulations of generic Trotterised circuits. In the first set of experiments, we vary the physical error rate from $10^{-6}$ to $10^{-2}$ for NISQ, partially fault-tolerant, and fully fault-tolerant circuits. Noise from high-weight Pauli rotation gates is modelled using global depolarising noise, and each logical gate layer is assumed to involve the same number of error correction cycles with the same associated logical noise. As shown in Fig. 2 (b), the TVD bound increases most rapidly for NISQ circuits, followed by partially fault-tolerant and then fully fault-tolerant circuits. As physical error rates approach the surface code threshold, the bounds for fault-tolerant circuits saturate near 1.

We next fix the physical error rate at $10^{-3}$ and increase the circuit depth up to 100 rotation layers. Fig. 2 (d) shows that the TVD bound again grows fastest for NISQ circuits, with fault-tolerant circuits degrading more slowly. We also examine how varying the surface code distance from 3 to 13 affects certification (Fig. 7). Fully fault-tolerant circuits benefit more from increased code distance, with sharper reductions in TVD at lower physical error rates.

Finally, Fig. 8 (b) shows the effect of magic state fidelity on circuit performance. As in the IQP case, reducing $T$-gate noise tightens the TVD bound until Clifford noise becomes dominant and the improvement plateaus.

To certify the full Hamiltonian simulation accuracy, logical accreditation must be combined with bounds on Trotterisation error. For second-order Trotterisation, this error scales as $Wt^3$, where $W$ is a Hamiltonian-dependent constant [45–47].

## V. APPLICATIONS

### A. Entropy density benchmarking for fault-tolerant computation

The method of *entropy density benchmarking* was developed to assess the quality of NISQ circuits in terms of entropy accumulation from circuit noise [17]. Recent work has extended this approach to create heuristic models of entropy accumulation in NISQ circuits run on currently available quantum hardware, using these to compute circuit-size thresholds beyond which quantum advantage is unattainable [18].

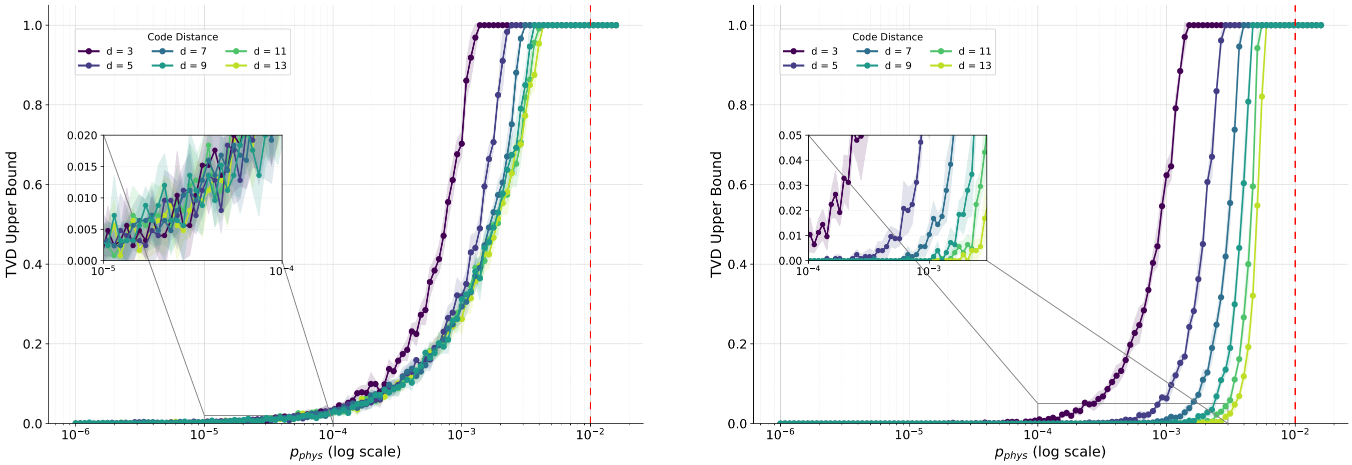
The second-order Rényi entropy density of the $n$ logical qubit $i$-th logical output state, $\rho_{out}^{(i)}$, is defined

$$n^{-1}S^{(2)}(\rho_{out}^{(i)}) := -n^{-1}\log_2(\mathrm{Tr}[\rho_{out}^{(i)}{}^2]), \tag{11}$$

where $n$ denotes the number of logical qubits used for the computation and $S^{(2)}(.)$ the second-order Rényi entropy. The $\gamma$ value computed during the logical accreditation protocol may be used to derive the following upper-bound on the entropy density of the target circuit logical output state

$$n^{-1}S^{(2)}(\rho_{out}^{(i)}) \leq -n^{-1}\log_2(1 - 2\gamma + \gamma^2(1 + 2^{-n})). \tag{12}$$

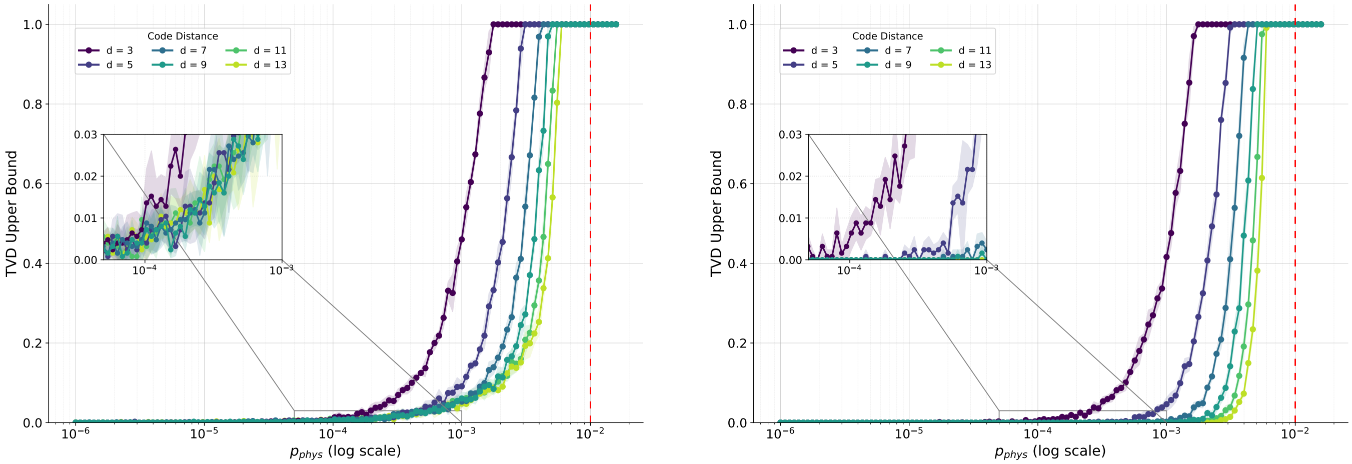
This bound, derived in Appendix H, holds with the same confidence as the TVD bound provided by the logical accreditation protocol. The logical accreditation protocol can, therefore, be used to efficiently compute an upper-bound on the Rényi entropy density of the target circuit, extending the entropy density benchmarking method to circuits run with encoded logical qubits. It would be interesting to further develop this connection, perhaps using logical accreditation as a means of creating heuristic models for entropy accumulation in logical circuits, similar to the approach followed in [18] for NISQ circuits.



(a) Partial fault tolerance

(b) Full fault tolerance

FIG. 6: *Impact of code distance on the TVD bound.* Simulations of 15-qubit, 40-layer IQP circuits are run using (a) partial and (b) full fault tolerance. Full fault tolerance shows greater sensitivity to distance and faster TVD convergence as the physical error rate decreases. The vertical dashed line indicates the surface code threshold.



(a) Partial fault tolerance

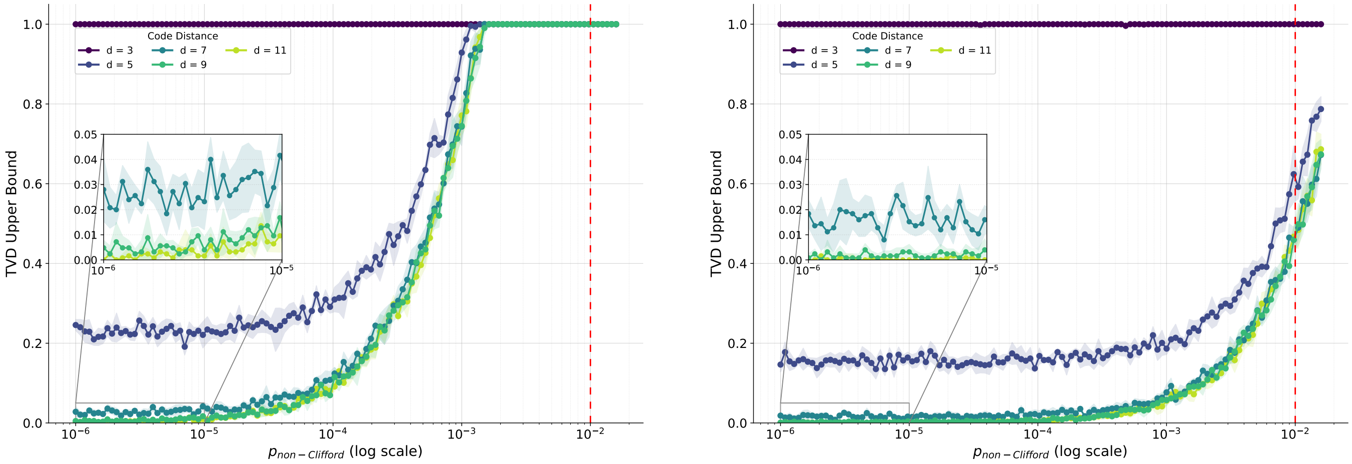
(b) Full fault tolerance

FIG. 7: *Effect of code distance on the certified TVD bound.* Simulations are run for 15-qubit, 40-layer IQP circuits using (a) partial and (b) full fault tolerance. Full fault tolerance shows greater sensitivity to distance and faster convergence of the TVD bound as physical error rate decreases and code distance increases. The vertical dashed line marks the surface code threshold.

### B. Certifying efficiency of quantum error mitigation applied to logical circuits

There has been much recent work focused on combining quantum error mitigation with quantum error correction for logical computations performed using early fault-tolerant quantum devices [1, 48–50]. Error mitigation techniques have been proposed as a means of extending the computational performance of logical computations with unpurified magic states [1], and where code distances are too low to suppress noise sufficiently to achieve desired logical error rates [49]. It has been asserted that quantum error mitigation is primarily useful in the regime $\epsilon_G N_G = O(1)$, where $\epsilon_G$ is the independent gate error rate and $N_G$ is the number of gates in the circuit [51, 52]. In previous work concerning the mitigation of gate errors for noisy physical qubits, the overall amplification in the standard deviation of the mitigated output was computed to be $(1+2\epsilon_G)^{2N_G} \sim e^{4N_G\epsilon_G}$ [53]. If the value of $N_G\epsilon_G$ is too large, error mitigation cannot practically be applied due to the exponential scaling of the sampling overhead. Similar arguments can be made concerning the efficiency of applying error mitigation to computations run on encoded logical qubits, where instead $\epsilon_G$ is the logical gate error rate and $N_G$ is the number of logical gates in the circuit.

Since logical accreditation provides a bound on the total circuit error rate, it can be used to bound the value of $N_G\epsilon_G$. For example, if the maximum acceptable overhead to apply error mitigation to a logical circuit is a fac-



(a) IQP circuits

(b) Trotterised circuits

FIG. 8: *Effect of magic state quality on certified performance across code distances.* We simulate 50-qubit, 40-layer IQP circuits (left) and Trotterised circuits (right), fixing the physical error rate at $p_{\text{phys}} = 10^{-3}$. The Clifford error rate is determined by the surface code distance, while the magic state error rate is varied. At low distances, the TVD bound does not vanish even as magic state fidelity improves, due to residual Clifford errors.

tor of 30 increase in the required number of samples relative to the unmitigated estimator, this condition is satisfied when $N_G \epsilon_G \leq 0.8503$. The probability of no errors occurring due to gate noise for a circuit with $N_G$ gates is $(1 - \epsilon_G)^{N_G}$, which is approximately $e^{-\epsilon_G N_G}$. Assuming errors only occur during the circuit due to gate noise, then the total circuit error rate is $p_{err} \approx 1 - e^{-\epsilon_G N_G}$. Setting $\epsilon_G N_G = 0.8503$ results in an upper-bound on the circuit error rate where mitigation can be efficiently applied of $p_{err} \leq 1 - e^{-0.8503}$. Since the $\gamma$ value provided by logical accreditation upper bounds both the TVD and the total circuit error rate, it can be used to check whether the error mitigation efficiency condition is satisfied. This follows as a corollary if the inequality $\gamma \leq 1 - e^{-0.8503}$ is experimentally satisfied. In this manner, logical accreditation can be used to certify the efficiency of applying quantum error mitigation to a logical circuit.

## VI. DISCUSSION AND CONCLUSION

In this work, we introduced logical accreditation as a framework for efficiently certifying fault-tolerant quantum computations. The protocol runs trap computations alongside the target computation, and analysis of the trap computation outcomes allows the accuracy of the target computation to be certified. Unlike traditional quantum error correction analyses, this approach is sensitive to a much broader class of noise models.

We demonstrated the protocol through numerical simulations of IQP circuit sampling and Trotterised quantum circuits. We also outlined some potential applications, including: (1) extending entropy density benchmarking to the fault-tolerant regime, and (2) assessing whether quantum error mitigation techniques can be efficiently applied to specific logical circuits. In addition, we introduced a compilation strategy that transforms general decoding and magic-state preparation noise into logical stochastic Pauli noise. As part of this, we proposed a method for twirling arbitrarily rotated magic states, showing that non-transversal logical gates, including those beyond the $T$ gate, can also be twirled.

There are many opportunities to extend this work. For example, the protocol could be optimised for specific quantum error-correcting codes, especially where code structure affects trap circuit design. It could also be run on near-term hardware to test the building blocks of fault tolerance and, as devices improve, to certify large fault-tolerant computations. The protocol could also be used to assess whether logical error rates are low enough that classical simulation algorithms, such as the one presented in [54], are inefficient.

### Acknowledgments

We would like to thank Andrew Jackson and Raúl García-Patrón for reading earlier versions of this manuscript and providing helpful feedback for its improvement, and also for interesting discussions. All authors acknowledge support from the EPSRC Quantum Advantage Pathfinder research program within the UK's National Quantum Computing Center. D.L. acknowledges support by the Autonomous Quantum Technologies (AutoQT) project (UKRI project 10004359). A.S. was supported by the Engineering and Physical Sciences Research Council (grant number EP/W524384/1). J. R. is funded by an EPSRC Quantum Career Acceleration Fellowship (grant code: UKRI1224). J. R. further acknowledges support from EPSRC grants EP/T001062/1 and EP/X026167/1.

[1] C. Piveteau, D. Sutter, S. Bravyi, J. M. Gambetta, and K. Temme, Physical Review Letters **127**, 200505 (2021), publisher: American Physical Society.

[2] R. Acharya et al., Nature , 1 (2024), publisher: Nature Publishing Group.

[3] D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, S. Ebadi, M. Cain, M. Kalinowski, D. Hangleiter, J. P. Bonilla Ataides, N. Maskara, I. Cong, X. Gao, P. Sales Rodriguez, T. Karolyshyn, G. Semeghini, M. J. Gullans, M. Greiner, V. Vuletić, and M. D. Lukin, Nature **626**, 58 (2024), publisher: Nature Publishing Group.

[4] A. Paetznick, M. P. d. Silva, C. Ryan-Anderson, J. M. Bello-Rivas, J. P. C. III, A. Chernoguzov, J. M. Dreiling, C. Foltz, F. Frachon, J. P. Gaebler, T. M. Gatterman, L. Grans-Samuelsson, D. Gresh, D. Hayes, N. Hewitt, C. Holliman, C. V. Horst, J. Johansen, D. Lucchetti, Y. Matsuoka, M. Mills, S. A. Moses, B. Neyenhuis, A. Paz, J. Pino, P. Siegfried, A. Sundaram, D. Tom, S. J. Wernli, M. Zanner, R. P. Stutz, and K. M. Svore, "Demonstration of logical qubits and repeated error correction with better-than-physical error rates," (2024), arXiv:2404.02280 [quant-ph].

[5] A. Katabarwa, K. Gratsea, A. Caesura, and P. D. Johnson, PRX Quantum **5**, 020101 (2024), publisher: American Physical Society.

[6] Y. Suzuki, S. Endo, K. Fujii, and Y. Tokunaga, PRX Quantum **3**, 010345 (2022), publisher: American Physical Society.

[7] A. Gheorghiu, T. Kapourniotis, and E. Kashefi, Theory of Computing Systems **63**, 715 (2019).

[8] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Physical Review A **77**, 012307 (2008), publisher: American Physical Society.

[9] E. Magesan, J. M. Gambetta, and J. Emerson, Physical Review Letters **106**, 180504 (2011), publisher: American Physical Society.

[10] J. Helsen, I. Roth, E. Onorati, A. Werner, and J. Eisert, PRX Quantum **3**, 020357 (2022), publisher: American Physical Society.

[11] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, Physical Review A **100**, 032328 (2019), publisher: American Physical Society.

[12] P. Jurcevic, A. Javadi-Abhari, L. S. Bishop, I. Lauer, D. F. Bogorin, M. Brink, L. Capelluto, O. Günlük, T. Itoko, N. Kanazawa, A. Kandala, G. A. Keefe, K. Krsulich, W. Landers, E. P. Lewandowski, D. T. McClure, G. Nannicini, A. Narasgond, H. M. Nayfeh, E. Pritchett, M. B. Rothwell, S. Srinivasan, N. Sundaresan, C. Wang, K. X. Wei, C. J. Wood, J.-B. Yau, E. J. Zhang, O. E. Dial, J. M. Chow, and J. M. Gambetta, Quantum Science and Technology **6**, 025020 (2021), publisher: IOP Publishing.

[13] E. Pelofske, A. Bärtschi, and S. Eidenbenz, IEEE Transactions on Quantum Engineering **3**, 1 (2022).

[14] J. J. Wallman and J. Emerson, Physical Review A **94**, 052325 (2016), publisher: American Physical Society.

[15] A. Hashim, R. K. Naik, A. Morvan, J.-L. Ville, B. Mitchell, J. M. Kreikebaum, M. Davis, E. Smith, C. Iancu, K. P. O'Brien, I. Hincks, J. J. Wallman, J. Emerson, and I. Siddiqi, Physical Review X **11**, 041039 (2021), publisher: American Physical Society.

[16] S. Ferracin, S. T. Merkel, D. McKay, and A. Datta, Physical Review A **104**, 042603 (2021), publisher: American Physical Society.

[17] D. Stilck França and R. García-Patrón, Nature Physics **17**, 1221 (2021), number: 11 Publisher: Nature Publishing Group.

[18] M. Demarty, J. Mills, K. Hammam, and R. Garcia-Patron, "Entropy Density Benchmarking of Near-Term Quantum Circuits," (2024), arXiv:2412.18007 [quant-ph].

[19] S. Ferracin, T. Kapourniotis, and A. Datta, New Journal of Physics **21**, 113038 (2019), publisher: IOP Publishing.

[20] A. Jackson, T. Kapourniotis, and A. Datta, Proceedings of the National Academy of Sciences **121**, e2309627121 (2024), publisher: Proceedings of the National Academy of Sciences.

[21] A. Jackson and A. Datta, "Improved Accreditation of Analogue Quantum Simulation and Establishing Quantum Advantage," (2025), arXiv:2502.06463 [quant-ph].

[22] J. F. Fitzsimons and E. Kashefi, Physical Review A **96**, 012303 (2017), publisher: American Physical Society.

[23] D. Leichtle, L. Music, E. Kashefi, and H. Ollivier, PRX Quantum **2**, 040302 (2021), publisher: American Physical Society.

[24] B. Eastin and E. Knill, Physical Review Letters **102**, 110502 (2009), publisher: American Physical Society.

[25] D. Horsman, A. G. Fowler, S. Devitt, and R. V. Meter, New Journal of Physics **14**, 123011 (2012), publisher: IOP Publishing.

[26] Y. Akahoshi, K. Maruyama, H. Oshima, S. Sato, and K. Fujii, PRX Quantum **5**, 010337 (2024), publisher: American Physical Society.

[27] Y. Akahoshi, R. Toshio, J. Fujisaki, H. Oshima, S. Sato, and K. Fujii, "Compilation of Trotter-Based Time Evolution for Partially Fault-Tolerant Quantum Computing Architecture," (2024), arXiv:2408.14929 [quant-ph].

[28] D. Litinski, Quantum **3**, 128 (2019), publisher: Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften.

[29] R. Toshio, Y. Akahoshi, J. Fujisaki, H. Oshima, S. Sato, and K. Fujii, "Practical quantum advantage on partially fault-tolerant quantum computer," (2024), arXiv:2408.14848.

[30] A. Y. Kitaev, Russian Mathematical Surveys **52**, 1191 (1997), publisher: IOP Publishing.

[31] A. Kitaev, A. Shen, and M. Vyalyi, "Classical and Quantum Computation," (2002), iSBN: 9780821832295 9781470409272 9781470420116 9781470418007 ISSN: 1065-7339 Publisher: American Mathematical Society Series: Graduate Studies in Mathematics Volume: 47.

[32] M. Suzuki, Journal of Mathematical Physics **32**, 400 (1991).

[33] D. W. Berry, G. Ahokas, R. Cleve, and B. C. Sanders, Communications in Mathematical Physics **270**, 359 (2007).

[34] D. Poulin, M. B. Hastings, D. Wecker, N. Wiebe, A. C. Doherty, and M. Troyer, "The Trotter Step Size Required for Accurate Quantum Simulation of Quantum Chemistry," (2014), arXiv:1406.4920 [quant-ph].

[35] S. Bravyi, M. Englbrecht, R. König, and N. Peard, npj Quantum Information **4**, 1 (2018), number: 1 Publisher: Nature Publishing Group.

[36] K. Tsubouchi, Y. Mitsuhashi, K. Sharma, and N. Yoshioka, "Symmetric Clifford twirling for cost-optimal quantum error mitigation in early FTQC regime," (2025), arXiv:2405.07720 [quant-ph].

[37] D. Shepherd and M. J. Bremner, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **465**, 1413 (2009), publisher: Royal Society.

[38] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Quantum **1**, 8 (2017), publisher: Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften.

[39] D. Hangleiter, M. Kalinowski, D. Bluvstein, M. Cain, N. Maskara, X. Gao, A. Kubica, M. D. Lukin, and M. J. Gullans, "Fault-tolerant compiling of classically hard IQP circuits on hypercubes," (2024), arXiv:2404.19005.

[40] M. E. Beverland, P. Murali, M. Troyer, K. M. Svore, T. Hoefler, V. Kliuchnikov, G. H. Low, M. Soeken, A. Sundaram, and A. Vaschillo, "Assessing requirements to scale to practical quantum advantage," (2022), arXiv:2211.07629 [quant-ph].

[41] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Physical Review Letters **117**, 080501 (2016), publisher: American Physical Society.

[42] S. Bravyi and D. Gosset, Physical Review Letters **116**, 250501 (2016), publisher: American Physical Society.

[43] H. Pashayan, O. Reardon-Smith, K. Korzekwa, and S. D. Bartlett, PRX Quantum **3**, 020361 (2022), publisher: American Physical Society.

[44] R. Nandkishore and D. A. Huse, Annual Review of Condensed Matter Physics **6**, 15 (2015), publisher: Annual Reviews.

[45] A. M. Childs, Y. Su, M. C. Tran, N. Wiebe, and S. Zhu, Physical Review X **11**, 011020 (2021), publisher: American Physical Society.

[46] I. D. Kivlichan, C. Gidney, D. W. Berry, N. Wiebe, J. McClean, W. Sun, Z. Jiang, N. Rubin, A. Fowler, A. Aspuru-Guzik, H. Neven, and R. Babbush, Quantum **4**, 296 (2020), arXiv:1902.10673 [quant-ph].

[47] E. T. Campbell, Quantum Science and Technology **7**, 015007 (2021), publisher: IOP Publishing.

[48] M. Lostaglio and A. Ciani, Physical Review Letters **127**, 200506 (2021), publisher: American Physical Society.

[49] Y. Suzuki, S. Endo, K. Fujii, and Y. Tokunaga, PRX Quantum **3**, 010345 (2022), publisher: American Physical Society.

[50] A. Dutkiewicz, S. Polla, M. Scheurer, C. Gogolin, W. J. Huggins, and T. E. O'Brien, "Error mitigation and circuit division for early fault-tolerant quantum phase estimation," (2024), arXiv:2410.05369 [quant-ph].

[51] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, Journal of the Physical Society of Japan **90**, 032001 (2021), publisher: The Physical Society of Japan.

[52] Z. Zimborás, B. Koczor, Z. Holmes, E.-M. Borrelli, A. Gilyén, H.-Y. Huang, Z. Cai, A. Acín, L. Aolita, L. Banchi, F. G. S. L. Brandão, D. Cavalcanti, T. Cubitt, S. N. Filippov, G. García-Pérez, J. Goold, O. Kálmán, E. Kyoseva, M. A. C. Rossi, B. Sokolov, I. Tavernelli, and S. Maniscalco, "Myths around quantum computation before full fault tolerance: What no-go theorems rule out and what they don't," (2025), arXiv:2501.05694 [quant-ph].

[53] S. Endo, S. C. Benjamin, and Y. Li, Physical Review X **8**, 031027 (2018), publisher: American Physical Society.

[54] T. Schuster, C. Yin, X. Gao, and N. Y. Yao, "A polynomial-time classical algorithm for noisy quantum circuits," (2024), arXiv:2407.12768.

[55] S. Bravyi and A. Kitaev, Physical Review A **71**, 022316 (2005), publisher: American Physical Society.

[56] Z. Liu, Y. Xiao, and Z. Cai, "Non-Markovian Noise Suppression Simplified through Channel Representation," (2024), arXiv:2412.11220 [quant-ph].

[57] A. Winick, J. J. Wallman, D. Dahlen, I. Hincks, E. Ospadov, and J. Emerson, "Concepts and conditions for error suppression through randomized compiling," (2022), arXiv:2212.07500 [quant-ph].

[58] P. Jordan and E. Wigner, Zeitschrift für Physik **47**, 631 (1928).

[59] R. Somma, G. Ortiz, J. E. Gubernatis, E. Knill, and R. Laflamme, Physical Review A **65**, 042323 (2002), publisher: American Physical Society.

# APPENDIX

## Appendix A: Trap circuit construction

The trap construction now described is the one used in the numerics, and in the derivation of the $\beta = 0$ and $\beta = 1/2$ upper bounds on the trap circuit error cancellation probability. In Appendix E 3, a modified trap construction is described that allows for $\beta = 7/8$ error cancellation probability upper bounds, this construction takes into account all possible error cancellation events.

The logic of each of the operations performed during the trap circuits either compiles to an identity operation or to an operation that stabilises the input state. This ensures that the trap circuits provide a deterministic bit string output in the absence of errors. The trap circuits are generated using the same circuit structure as the target circuits. The positioning and type of the multi-qubit Clifford gates are kept the same. The single-qubit Clifford gate layers are replaced with randomly chosen $S$, $S^\dagger$ and $H$ gates. When randomly applied $S$, $S^\dagger$ and $H$ operations sandwich a C$Z$ gate, the resulting combined operation is logically equivalent to a randomly oriented CNOT gate, as shown in Fig. 11.

A layer of logical Hadamard gates is included at the beginning and end of the trap circuits with probability 1/2. An important function of the randomly chosen $S$ and $H$ operations is to randomly map logical Pauli error operators to different logical Pauli operators, preventing the same types of error cancellation from happening in different trap computations. This property is used to derive the protocol guarantees in Appendix D.

We now describe the treatment of gates performed by the consumption of both unpurified and purified magic states.

### 1. Logical gates performed using magic states without magic state purification

In the trap circuits, gates corresponding to logical single- and multi-qubit Pauli rotation gates in the target circuit are modified to stabilise the logical state. The magic state preparation procedure is different for the trap and target circuits. We will refer to magic states used in the target circuit as target magic states, and magic states used in the trap circuits as trap magic states.

In the scenario where target and trap magic states are not purified before use, two versions of each trap circuit are run on the quantum device. If at least one of the two versions of the trap circuit detects an error, this is recorded as a failure for that trap. In one version, all magic states prepared are $|\pi\rangle$ states; in the other, instead $|\pi/2\rangle$ states are prepared, each followed by a fault-tolerant $S$ gate. Although the state preparation is performed differently, both versions ultimately produce $|\pi\rangle$ states. The reason for this approach is that $|\pi\rangle$ magic states are only vulnerable to $Y$ and $Z$ Pauli errors during state preparation, whereas $|\pi/2\rangle$ magic states are vulnerable to $X$ and $Z$ Pauli errors. If a trap magic state is used to perform a Pauli rotation gate, a single fault-tolerant Pauli $\pi$-rotation gate is applied immediately after the gate is applied. In the absence of errors, the combined logic of these two operations is an identity operation. In practice, the Pauli correction gates may be absorbed into the following logical randomised compiling Pauli gate layer.

The target and trap magic states are prepared using protocols in which the phase rotation of the prepared states is determined by physical single-qubit Pauli rotation gates after the logical $|+\rangle$ state has been encoded. A consequence of assumption (A1), that single-qubit gate noise is gate independent, is that magic state preparation noise is phase rotation-angle independent. Two examples of logical gate operations for magic state preparation that apply an arbitrary logical phase rotation to a $|+\rangle$ state are shown in Fig. 9 (b) and (c). These operations are from the partially fault-tolerant STAR framework [26], and apply a logical $R_Z(\theta)$ operation using a single physical single-qubit Pauli rotation gate, along with C$Z$ and CNOT gates. And so, under the assumption that physical single-qubit gate noise is gate-independent, the logical state-preparation noise for the $|\pi/4\rangle$ and $|\theta\rangle$ target magic states, and the $|\pi\rangle$ and $|\pi/2\rangle$ trap magic states, is the same.

Each gate applied using a magic state is randomly sandwiched by $S$, $S^\dagger$ and $H$ gates. For RUS gates, each gate repetition is sandwiched by these operations. The random sandwiching gate operations are used to randomise error propagation through the trap circuits, ensuring errors are detected with constant probability.

### 2. Logical gates performed using purified magic states

We now describe how magic state purification with the Clifford+$T$ framework can be incorporated into the certification protocol. To perform fully fault-tolerant computation with a universal gate set, purification must be used to improve magic state fidelities. Purification can be used to ensure that gates performed by the consumption of magic states have error rates similar to those of logical gates protected by the QEC code. For the protocol to certify logical computations with purified magic states, trap circuits must be constructed that incorporate trap magic state purification. Purified trap magic states must have error rates similar to the purified target magic states for the output of the certification protocol to be valid. The most well-known magic state purification protocol is magic-state distillation [55]. During distillation, a number of noisy input encoded $|\pi/4\rangle$ magic states are purified, or distilled, to generate a smaller number of higher-quality states. This is usually done by concatenating the code used to encode the computational logical qubits with another code possessing a transversal $T$ gate. The concatenated code is used for error detection, postselecting on instances where no error is detected in order to generate higher quality magic states. For the distillation to be successful, the fidelities of the input magic states must be greater than the fidelity threshold of the distillation method used. If distillation is successfully repeated, at each repetition, magic states of progressively higher quality are generated. When a magic state of sufficiently high quality is produced it is then used to perform a $\pi/4$ rotation gate on the computational logical qubits. Fully fault-tolerant computation is possible if purified magic states are used to perform those gates not protected by the QEC code.

To avoid making additional assumptions about the similarity of the purified target and trap magic state fidelities, $|\pi/4\rangle$ states are purified for both the target and trap circuits. If there are $K$ gates in the target or trap circuit performed using a magic state, $2K$ purifed $|\pi/4\rangle$ states are initially prepared, and two magic states are assigned to each gate. For each gate in the target computation, one $|\pi/4\rangle$ state is discarded and the other is kept and used to perform the gate. For each gate in a trap computation, two purified $|\pi/4\rangle$ states are combined into a $|\pi/2\rangle$ state, and a fault-tolerant $S$ gate is then applied to make it a $|\pi\rangle$ state, this is then used to perform the gate. Circuit diagrams depicting these operations are shown in Fig. 10.

In some cases, it may be reasonable to assume that purification of $|\pi/2\rangle$ or $|\pi\rangle$ states provides output states of very similar quality to the output states obtained from $|\pi/4\rangle$ state purification. This would depend on the type of purification method used and the form of the logical noise affecting the purification circuits. The reason this assumption is avoided in the previously stated approach is that $|\pi/4\rangle$ states are vulnerable to $X$, $Y$ and $Z$ er-

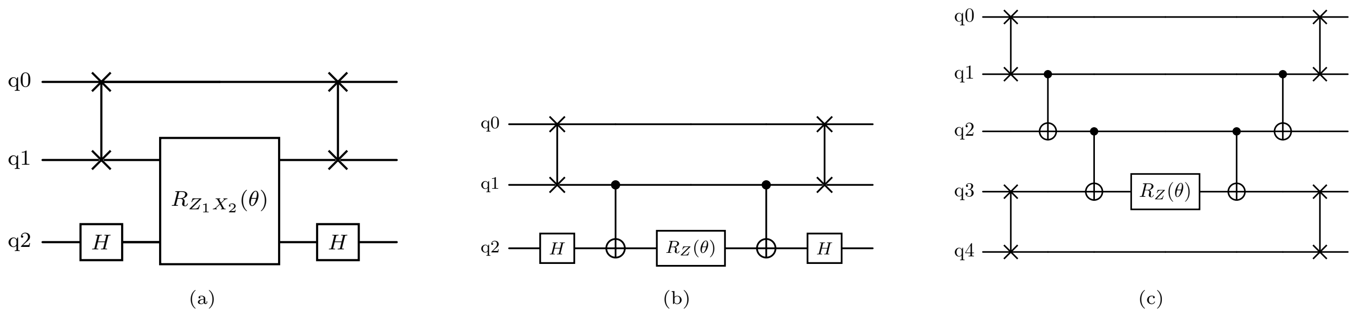


FIG. 9: *Examples of multi-qubit Z rotations used in magic state preparation in the space-time efficient analog rotation scheme [26, 29].* In (a) a weight-2 $ZZ$ rotation is applied to two physical qubits. While in (b) and (c) a local $Z$ rotation is applied to a single qubit, that is then compiled through the use of additional CNOT operations into a $ZZ$ and a $ZZZ$ rotation, respectively. A requirement of the certification protocol is that the magic state rotation angles be determined by the single application of a local rotation gate. Therefore, the circuit shown in (a) is incompatible with the protocol, whereas the circuits shown in (b) and (c) are compatible with the protocol.

rors, while $|\pi/2\rangle$ states are stabilised by $Y$ errors and $|\pi\rangle$ states are stabilised by $X$ errors. As a result, purifying different states for target and trap circuits could lead to appreciable differences in the fidelities of the purified target and trap magic states, potentially affecting the validity of certification bound. However, if this assumption is accurate, one could directly purify the trap and target magic states and use these to perform gates, removing the need for additional magic states.

As in the case where no purification is applied, the gates performed by consumption of the purified trap magic states are immediately followed by a cancelling fault-tolerant Pauli $\pi$ rotation gate, resulting in overall identity operation. Also like before, these gates are sandwiched by randomising $S$, $S^\dagger$ and $H$ gates.

## Appendix B: Logical randomised compiling

We now describe logical randomised compiling, a method for compiling logical circuit that effectively transforms general logical circuit noise, from, for example, incorrect decoding or imperfect magic state preparation, into logical stochastic Pauli noise. This is based on the NISQ circuit compilation technique called randomised compiling, where randomly chosen Pauli gates are compiled together with layers of single-qubit gates to Pauli twirl circuit noise into stochastic Pauli noise [14]. We now describe how to compile the twirling operations into the logical circuits, and how these operations transform noise affecting different components of the logical circuit. This includes the twirling of magic state preparation noise, in the scenario where state purification is applied, and where it is not.

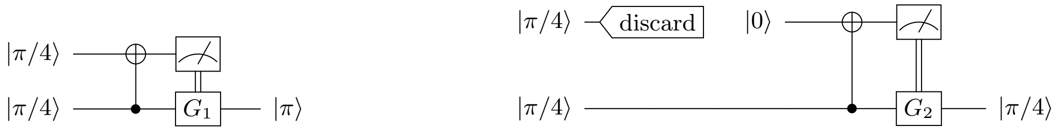
It was shown in [56] and [57] that performing Pauli twirling on non-Markovian noise channels generates effective stochastic Pauli channels, much like performing Pauli twirling on individual noise layers as if they are Markovian. Unlike twirling Markovian noise, however, Pauli twirling non-Markovian noise can generate effective stochastic Pauli channels that can be classically correlated in time. The logical randomised compiling procedure remains the same for Markovian and non-Markovian noise. The analysis in this section treats the logical noise as Markovian, but this naturally extends to the non-Markovian case in the same way as was shown for NISQ randomised compiling in [56] and [57]. If logical noise is non-Markovian, logical randomised compiling generates effective stochastic Pauli channels that can be correlated with an environment, and can cause error correlations in time. Specifically, the different logical stochastic Pauli channels affecting a circuit can have Pauli coefficients that are classically correlated in time. This has implications for the soundness of the certification protocol that are discussed in Appendix E.

### 1. Compiling twirling operations into circuits

Since the logical qubits are encoded using QEC code, logical single-qubit gates and logical Pauli gates cannot be compiled together, as is done in NISQ randomised compiling [14]. We now describe how the logical randomised compiling operations are integrated into the logical target and trap circuits. The twirling operations are layers of logical single-qubit Pauli operations interleaving all the logical gate layers of the circuit. Contiguous twirling operations are compiled together, so that a single logical layer of Pauli twirling operations includes both the undoing operation from the previous twirl and the operation for the following twirl. For example, a single instance of a twirling gate layer $P_i'$ performed at an arbitrary point in the circuit combines both the undoing operation for the $(i-1)$-th twirl, $P_i^{(1)}$ and the application of $i$-th twirl, $P_{i-1}^{(2)}$, that is

$$P_i' = P_i^{(1)} P_{i-1}^{(2)}. \tag{B1}$$

Then, rather than having a logical noise channel for both $P_i$ and $P_{i-1}^\dagger$, there is a single noise channel associated with the logical Pauli operation $P_i'$. So the noisy appli-



(a)

(b)

FIG. 10: *The final step of the magic state preparation where purification is used, where either a trap or target magic state is constructed from two purified states.* (a) To produce a trap magic state, gate teleportation is used between the two input purified $|\pi/4\rangle$ states to produce a $|\pi\rangle$ state. The measurement operation is performed in the $Z$ basis and if the measurement returns the value $+1$ the gate operation $G_1$ is an $S$ gate, while if the measurement returns $-1$, the gate operation $G_1$ is a $Z$ gate. (b) To produce a target magic state one of the input purified $|\pi/4\rangle$ states is immediately discarded and a $|0\rangle$ state initialised. The measurement is in the $Z$ basis, and if the measured value is $+1$, the gate operation $G_1$ applied is an $I$ gate, while if the measurement returns $-1$, the gate operation $G_1$ applied is an $S$ gate. This is a dummy gate teleportation procedure that is performed in order to make the operations that produce the purified target and trap magic states as similar as possible.

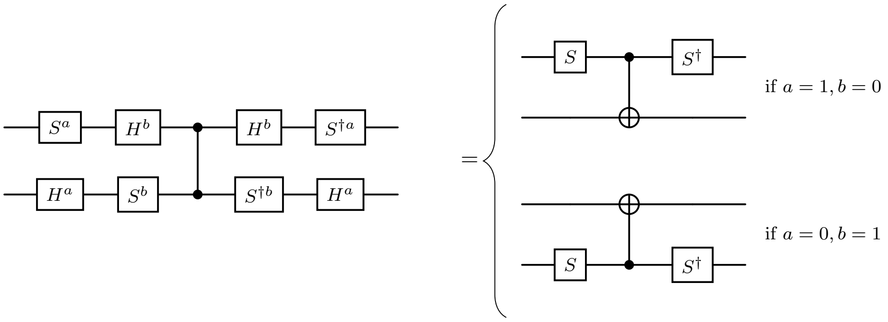


FIG. 11: In the trap circuits the C$Z$ gates are randomly sandwiched by $S$ and $H$ gates such that the combined logic of the gates is a randomly oriented CNOT gates by appending $S$ and $H$ gates. This process is shown in the diagrammatic circuit equation, where the bit $b \in \{0, 1\}$ is picked uniformly at random for each C$Z$ gate instance on the LHS, and $a = b \oplus 1 \mod 2$.

cation of $P_i'$ may be written

$$\tilde{P}_i' = \mathcal{E}_{P_i'} P_i^{(1)} P_{i-1}^{(2)}, \tag{B2}$$

where the noise channel $\mathcal{E}_{P_i'}$ is twirling gate layer independent (see Assumption (2) in Section II B) and so does not depend on the specific twirling operation being applied.

### a. Twirling logical computational state preparation noise

The initial logical state for the protocol is the $|0^n\rangle := |0\rangle^{\otimes n}$ state. If the prepared logical state is denoted by $\rho$, the logical state preparation error rate is defined as

$$\epsilon := 1 - \langle 0^n| \rho |0^n\rangle . \tag{B3}$$

The twirling gate set is chosen to be $\{I, Z\}^{\otimes n}$, and the twirled logical state may then be written

$$\begin{aligned}\rho_{\mathcal{T}} &= 2^{-n} \cdot \sum_{P_i \in \{I,Z\}^{\otimes n}} P_i \rho P_i \\ &= 2^{-n} \cdot \sum_{P_i \in \{I,Z\}^{\otimes n}} P_i \Bigg( \sum_{x \in \{0,1\}^n} \sum_{y \in \{0,1\}^n} \alpha_{x,y} |x\rangle \langle y| \Bigg) P_i .\end{aligned} \tag{B4}$$

For each $\alpha_{x,y} |x\rangle \langle y|$ term in the sum within the brackets, if $x \neq y$ then the term acquires a multiplicative prefactor of $-1$ from half the Pauli operators, and a prefactor of $+1$ from the other half. While if $x = y$ then the term will acquire a multiplicative prefactor of $+1$ from all the Pauli operators. In consequence, the off-diagonal terms cancel, leaving only the diagonal terms in the sum

$$\begin{aligned}\rho_{\mathcal{T}} = &\sum_{x \in \{0,1\}^n} \alpha_{x,x} |x\rangle \langle x| . \\ &= (1-\epsilon) |0^n\rangle \langle 0^n| + \epsilon \cdot \sum_{s \in \{0,1\}^n / 0^n} \alpha_{s,s} |s\rangle \langle s| \\ &= (1-\epsilon) |0^n\rangle \langle 0^n| \\ &\qquad + \epsilon \cdot \sum_{P_i \in \{I,X\}^{\otimes n} / I^{\otimes n}} \alpha_{P_i} P_i |0^n\rangle \langle 0^n| P_i ,\end{aligned} \tag{B5}$$

with $\sum_{s \in \{0,1\}^n / 0^n} \alpha_{s,s} = \sum_{P_i \in \{I,X\}^{\otimes n} / I^{\otimes n}} \alpha_{P_i} = 1$. And so the state preparation noise is twirled into a stochastic Pauli-$X$ channel.

### b. Twirling logical Clifford gates

Single and multi-qubit Clifford gates in the target and trap circuits may be efficiently Pauli twirled. The Pauli gates applying each twirl are chosen uniformly at random, and, since the gates being twirled are Clifford, the gates resulting from propagating the initial Pauli gate operations through them may be efficiently classically

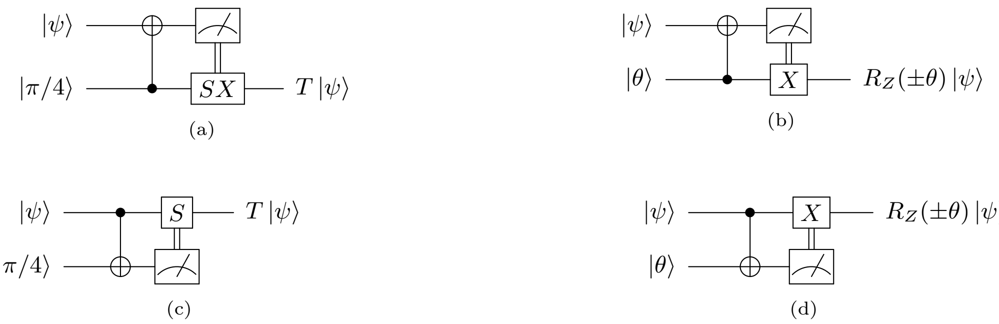


FIG. 12: Quantum circuit diagrams showing different ways of performing $T$ and RUS $R_Z(\theta)$ operations by gate teleportation including Pauli twirling gates.

computed. The twirling operations leave the logical circuit unchanged, only transforming general decoding errors that might occur during the application of the logical Clifford gates into logical stochastic Pauli errors. The twirling may be performed in a similar way to the randomised compiling technique proposed in [14], where physical Pauli operations are used to twirl noisy physical CNOT gates.

### *c. Twirling logical non-Clifford Pauli rotations*

An $n$-qubit logical Pauli rotation $R_P(\theta)$, where $P \in \{I,X,Y,Z\}^{\otimes n}$, may be performed using either gate teleportation or projective measurement, through the consumption of a $|\theta\rangle$ magic state. We will now show that twirling such a gate using the approach described transforms any logical Pauli rotation gate noise from imperfect magic state preparation, teleportation or projective measurement into a logical stochastic Pauli channel. We note that the method presented here provides one possible resolution to the open question posed by Pivateau et al., about how one might twirl the logical noise of non-transversal non-Clifford gates beyond the $T$ gate [1].

The binary symplectic representation of Pauli operators defined by $\phi(.)$ is the isomorphic map from the group of phaseless $n$-qubit Pauli operators to $2n$-bit binary vector, that is $\mathcal{P}_n \to \mathbb{Z}_2^{2n}$. Where the $j^{th}$ entry of the vector is equal to the power of the $X$ operator at the corresponding position in the Pauli operator, and the $(j+n)^{th}$ entry the power of the $Z$ operator in the corresponding position. The action of the map $\phi$ on an arbitrary Pauli $P \in \mathcal{P}_n$ is

$$\begin{aligned}\phi(P) &= \phi\big(\bigotimes_{j=1}^{n} X^{x_j} Z^{z_j}\big)\\ &= \big(\bigoplus_{j=1}^{n} x_j\big) \oplus \big(\bigoplus_{j=1}^{n} z_j\big)\\ &= (\mathbf{x}|\mathbf{z}).\end{aligned} \tag{B6}$$

Now the commutation relation of two Pauli operators, $P_1 = X^{\mathbf{x_1}} Z^{\mathbf{z_1}}$ and $P_2 = X^{\mathbf{x_2}} Z^{\mathbf{z_2}}$, may be written

$$(X^{\mathbf{x_1}} Z^{\mathbf{z_1}})(X^{\mathbf{x_2}} Z^{\mathbf{z_2}}) = (-1)^{\mathbf{x_1}\cdot\mathbf{z_2}+\mathbf{x_2}\cdot\mathbf{z_1}}(X^{\mathbf{x_2}} Z^{\mathbf{z_2}})(X^{\mathbf{x_1}} Z^{\mathbf{z_1}}) \tag{B7}$$

The term dictating the power of $(-1)$ defines the symplectic product, this operation preserves the commutation properties of the Pauli group in this binary representation. The symplectic product is defined

$$(\mathbf{x_1}|\mathbf{z_1}) \odot (\mathbf{x_2}|\mathbf{z_2}) := \mathbf{x_1}\cdot\mathbf{z_2} + \mathbf{x_2}\cdot\mathbf{z_1} \bmod 2, \tag{B8}$$

and two Pauli operators commute iff

$$(\mathbf{x_1}|\mathbf{z_1}) \odot (\mathbf{x_2}|\mathbf{z_2}) = 0, \tag{B9}$$

where the space $(\mathbb{F}_2^n \oplus \mathbb{F}_2^n, \odot)$ is sometimes referred to as a symplectic product space.

Let the initial $m$-qubit quantum state, where $m \geq n$, be denoted by $\rho$. And let the application of the noisy rotation gate be modelled as $\mathcal{E} \circ R_P(\theta) \circ (.)$, where $\mathcal{E}(.)$ is the logical noise channel associated with the gate. Applying the noisy Pauli rotation to the initial state $\rho$ then gives

$$\rho' = \mathcal{E} \circ R_P(\theta) \circ (\rho), \tag{B10}$$

Now including Pauli twirling operations for the noisy rotation gate results in the state

$$\begin{aligned}\rho_{\mathcal{T}} &= 4^{-n}\Bigg(\sum_{P_i\in\{I,X,Y,Z\}^{\otimes n}} P_i \circ \mathcal{E} \circ R_P(\theta) \circ P_i\Bigg) \circ (\rho)\\ &= 4^{-n}\Bigg(\sum_{P_i\in\{I,X,Y,Z\}^{\otimes n}} P_i \circ \mathcal{E} \circ P_i\Bigg)\\ &\qquad \circ R_P\big((-1)^{b_{P_i}\odot b_P}\cdot\theta\big) \circ (\rho),\end{aligned} \tag{B11}$$

where $b_{P_i}$ and $b_P$ are the binary symplectic representations of the Pauli operators $P_i$ and $P$ respectively. To get the second equality, the Pauli operators have been propagated through the $R_P(\theta)$ gate to isolate the logical error channel for the Pauli twirling. However, the Pauli operators that do not commute with $P$ pick up a phase of $-1$. This phase can be removed from the twirled state by updating the magic state preparation depending on the twirling gates applied in the following way.

We are considering the situation where magic state preparation is performed by applying a non-transversal logical Pauli-$Z$ rotation to a logical $|+\rangle$ state, where the phase rotation angle depends on the action of physical $Z$ rotation gates (e.g., Fig. 9 (b) and (c)). And so, assumption A1, that physical single-qubit gate noise is gate-independent, means that the rotation angle of the physical Pauli-$Z$ gate applied during magic state preparation may be changed without changing the logical gate noise. Therefore, the logical error channel $\mathcal{E}$ may be isolated for the Pauli twirling by reversing the rotation angle of the physical $R_P$ to $-\theta$ when Pauli twirling operations are applied that anticommute with $P$. So that, including the phase updates in the rotation angles, the twirled state is

$$\begin{aligned}\rho_{\mathcal{T}} = 4^{-n}\Bigg( &\sum_{P_i \in \{I,X,Y,Z\}^{\otimes n}} P_i \circ \mathcal{E} \circ \\ &R_P\big((-1)^{b_{P_i}\odot b_P}\cdot\theta\big) \circ P_i\Bigg) \circ (\rho) \\ = 4^{-n}\Bigg( &\sum_{P_i \in \{I,X,Y,Z\}^{\otimes n}} P_i \circ \mathcal{E} \circ P_i\Bigg) \\ &\circ R_P(\theta) \circ (\rho),\end{aligned} \tag{B12}$$

The twirled magic state with the angle updates is then

$$\rho'_{\mathcal{T}} = \Bigg(4^{-n}\cdot \sum_{P_i \in \{I,X,Y,Z\}^{\otimes n}} P_i \circ \mathcal{E} \circ P_i\Bigg) \circ R_P(\theta) \circ (\rho). \tag{B13}$$

The term inside the brackets is in the standard form required for Pauli twirling a quantum channel. Now suppose the channel $\mathcal{E}$ has Kraus operators $\{B_k\}_k$, that obey the trace-preserving property $\sum_k B_k^\dagger B_k = I$, known as the completeness condition. Channels that may be represented by a single Kraus operator are known as coherent, and when more than one Kraus operator is required, as incoherent. The evolution of state $\rho$ under the action of $\mathcal{E}$ may be modelled using the operator sum representation

$$\mathcal{E}(\rho) = \sum_k B_k \rho B_k^\dagger. \tag{B14}$$

The $n$-qubit Pauli basis is a complete basis for $n$-qubit quantum channels. Therefore, each Kraus operator in the set $\{B_k\}_k$ may be decomposed in the Pauli basis such that

$$B_k = \sum_i \gamma_i^k P_i, \tag{B15}$$

where $\{P_i\}_i$ is the set of all $n$-qubit Pauli operators and $|\{P_i\}_i| = 4^n$. Now the operator sum representation of the state evolution may be decomposed in the Pauli basis as

$$\begin{aligned}\mathcal{E}(\rho) &= \sum_k \Big(\sum_i \gamma_i^k P_i\Big)\rho\Big(\sum_j \gamma_j^k P_j\Big)^\dagger \\ &= \sum_k \sum_{i,j} \big(\gamma_i^k P_i\big)\rho\big(\gamma_j^k P_j\big)^\dagger \\ &= \sum_{i,j}\sum_k \gamma_i^k \gamma_j^{k*} P_i \rho P_j \\ &= \sum_{i,j} \chi_{i,j} P_i \rho P_j,\end{aligned} \tag{B16}$$

where in the final line we have defined $\chi_{i,j} = \sum_k \gamma_i^k \gamma_j^{k*}$. Indeed this is the well known process or $\chi$-representation, and the matrix defined by $\chi_{i,j}$ known as the $\chi$-matrix. The relation $\sum_{i,j} \chi_{i,j} P_j P_i = I$ follows from the trace preserving property of the set of Kraus operators representing a quantum channel. From which it follows that $\mathrm{tr}(\sum_{i,j} \chi_{i,j} P_j P_i) = 1$, or, equivalently,

$$\begin{aligned}\sum_i \chi_{i,i} &= \sum_{i,k} \gamma_i^k \gamma_i^{k*} \\ &= \sum_{i,k} |\gamma_i^k|^2 \\ &= 1,\end{aligned} \tag{B17}$$

therefore the diagonal entries of the chi matrix are real and sum to 1.

Now

$$\begin{aligned}&4^{-n}\sum_l P_l P P_l \rho P_l P' P_l \\ &\quad 4^{-n}\sum_l (-1)^{b_{P_l}\odot b_P} P \rho (-1)^{b_{P_l}\odot b_{P'}} P' \\ &\quad 4^{-n}\sum_l (-1)^{b_{P_l}\odot b_P + b_{P_l}\odot b_{P'}} P \rho P' \\ &\quad 4^{-n}\sum_l (-1)^{b_{P_l}\odot (b_P + b_{P'})} P \rho P'.\end{aligned} \tag{B18}$$

The distributivity of the symplectic inner product is used in the final equality. Now if $P = P'$, then the symplectic product becomes $b_{P_l} \odot 0 = 0$ and then result is

$$4^{-n}\sum_l P_l P P_l \rho P_l P P_l = P \rho P, \tag{B19}$$

Whereas if $P \neq P'$, then $(b_P + b_{P'})$ for half of the Pauli group $b_{P_l} \odot (b_P + b_{P'}) = 0$ and half $b_{P_l} \odot (b_P + b_{P'}) = 1$ resulting in overall cancellation and the elimination of the term. So that

$$4^{-n}\sum_l P_l P P_l \rho P_l P' P_l = 0 \quad \forall P \neq P'. \tag{B20}$$

An intuitive interpretation of this is that, if $P \neq P'$, the sum mod 2 of two binary representations $b_P + b_{P'}$

defines a new $n$-qubit Pauli operator, and this operator commutes with half of the Pauli group and anticommutes with the other half, resulting in cancellation.

The Pauli twirl eliminates all of the off-diagonal components of the Pauli decomposition of the channel, leaving only the diagonal elements. As these elements are real, and sum to 1, the transformed channel is a stochastic Pauli channel. Taking the Pauli twirl of the channel $\mathcal{E}$ transforms it to

$$\begin{aligned}\mathcal{E}'(\rho) &= |\{P_l\}_l|^{-1}\sum_l\sum_{i,j}\chi_{i,j}P_lP_iP_l\rho P_lP_jP_l \\ &= \sum_i \chi_{i,i}P_i\rho P_i \\ &= \sum_i c_{P_i}P_i\rho P_i.\end{aligned} \tag{B21}$$

Where we have defined the set of Pauli operator probability coefficients $\{c_{P_i}\}_i$, such that the coefficient for Pauli operator $P_i$ is $c_{P_i}$. Therefore, the magic state preparation noise is transformed by the twirling into stochastic Pauli noise.

### *d. Logical measurement twirling*

Logical measurement noise may be Pauli twirled by applying random Pauli operations immediately prior to logical measurement, and then undoing the action of the Paulis after measurement through the use of classical postprocessing. Noise affecting the measurement outcome $|z\rangle$, where $z \in \{0,1\}^n$, is modelled as $\mathcal{E}_M \circ (|z\rangle\langle z|)$ where the logical measurement noise is denoted by $\mathcal{E}_M$. The measurement noise may then be Pauli twirled by applying a layer of logical Pauli operations immediately prior to measurement, and a virtual Pauli gate layer after measurement between the noise channel $\mathcal{E}_M$ and the ideal measurement outcome $|z\rangle\langle z|$. The virtual gate is performed by classical postprocessing the measured output bit string, conditionally flipping the output bits, to undo the effect of the logical Pauli layer on the measurement outcome. That is, if the $i$-th term in the tensor product of the Pauli operator applied before measurement is an $X$ or a $Y$ operator, the $i$-th bit in the output bit string is flipped. Whereas if it is an $I$ or a $Z$ operator, the bit is left unchanged. This may be summarised by the relation $P_i|z\rangle\langle z|P_i = |z \oplus x_{P_i}\rangle\langle z\oplus x_{P_i}|$, where the sum is mod 2 and the bits of the ordered bit string $x_{P_i}$ are 1s where the logical $n$-qubit Pauli operator $P_i$ acts on the corresponding logical qubit with an $X$ or $Y$ operator, and 0s otherwise. The measurement noise is then transformed into a logical stochastic Pauli channel in the following way.

The logical computational measurement error rate is defined as

$$\epsilon := 1 - \langle z|\left(\mathcal{E}_M \circ (|z\rangle\langle z|)\right)|z\rangle\,. \tag{B22}$$

The twirling operations are applied immediately before measurement, and the twirled measurement outcome may then be written

$$\begin{aligned}M_{\mathcal{T}} &= 4^{-n}\cdot\sum_{P_i\in\{I,X,Y,Z\}^{\otimes n}} P_i\circ\mathcal{E}_M\circ(|z\oplus x_{P_i}\rangle\langle z\oplus x_{P_i}|) \\ &= 4^{-n}\cdot\sum_{P_i\in\{I,X,Y,Z\}^{\otimes n}} P_i\circ\mathcal{E}_M\circ(P_i|z\rangle\langle z|P_i) \\ &= 4^{-n}\cdot\sum_{P_i\in\{I,X,Y,Z\}^{\otimes n}} P_i\circ\mathcal{E}_M\circ P_i\circ(|z\rangle\langle z|) \\ &= (1-\epsilon)|z\rangle\langle z| \\ &\quad + \epsilon\cdot\sum_{P_j\in\{I,X,Y,Z\}^{\otimes n}/I^{\otimes n}}\alpha_{P_j}P_j\circ(|z\rangle\langle z|).\end{aligned} \tag{B23}$$

To ensure normalisation $\sum_{P_j\in\{I,X,Y,Z\}^{\otimes n}/I^{\otimes n}}\alpha_{P_j} = 1$. The relation $P_i|z\rangle\langle z|P_i = |z\oplus x_{P_i}\rangle\langle z\oplus x_{P_i}|$ is used in the second line, where the notation $\oplus$ denotes addition modulo two. This operation can be achieved by reassigning bits in the measured outcomes.

### *e. Compiling the twirling gate operations into the logical circuits*

The Pauli twirling operations are included in the target and trap circuits by effectively interleaving gate layers with Pauli twirling gate layers, with the first Pauli layer of each circuit used to twirl state-preparation noise and the last to twirl measurement noise. The $i$-th logical circuit, $\mathcal{C}^{(i)}$, may be written as a sequence of $D$ logical gate layers and logical Pauli twirling gate layers

$$\tilde{\mathcal{C}}^{(i)\prime} = \mathcal{P}^{(1)}_{D+1}\bigcirc_{j=1}^{D}(\mathcal{P}^{(2)}_j\circ\mathcal{L}^{(i)}_j\circ\mathcal{P}^{(1)}_j)\circ\mathcal{P}^{(2)}_0, \tag{B24}$$

where $\mathcal{L}^{(i)}_j$ denotes the $j$-th logical gate layer of the $i$-th circuit. For the $j$-th Pauli twirl, the notation $\mathcal{P}^{(1)}_j$ and $\mathcal{P}^{(2)}_j$ indicates the Pauli twirling operation being applied and then undone, respectively. The $\mathcal{P}^{(1)}_0$ and $\mathcal{P}^{(1)}_{D+1}$ Pauli gate layers twirl the logical state preparation noise and the logical computational qubit measurement noise. Contiguous twirling operation layers may be compiled together into a single Pauli layer. The result of the Pauli twirling operations is that all logical state-preparation, gate and measurement noise becomes logical stochastic Pauli noise. The trap circuits are designed such that all of these logical Pauli noise channels circuit independent but depending on their positioning in the circuit, such that $\mathcal{E}^{(i)}_j = \mathcal{E}^{(k)}_j$ for all $i,k\in\{1,...,M+1\}$ and $j\in\{1,\ldots,M\}$.

## 2. Twirling magic state preparation noise

We distinguish between two frameworks for logical computation, describing means of twirling magic state preparation noise into logical stochastic Pauli noise in each instance. Firstly, in the case where target magic states are of the form $|\pi/4\rangle$, these states can be purified or unpurified. And secondly, in the case where target magic states are arbitrarily phase-rotated and are of the form $|\theta\rangle$, these states are not purified.

### ***Twirling scheme for computations involving $|\pi/4\rangle$ target magic states***

If $|\pi/4\rangle$ magic states are used in the target circuit, the following state twirling methods are used. This approach is compatible with the inclusion of magic state purification in the computation.

We now describe the procedures for twirling the state-preparation noise of the $|\pi/4\rangle$ target magic states, and the $|\pi/2\rangle$ and $|\pi\rangle$ trap magic states into stochastic logical Pauli noise. Although the proofs demonstrating that twirling the state-preparation noise of the $|\pi/2\rangle$ and $|\pi\rangle$ magic states results in stochastic $Z$-channels are similar to the one for $|\pi/4\rangle$ magic states, we include them for the sake of completeness.

The initial state in each instance is a logical $|+\rangle$ state; the state preparation noise may be twirled using the gate set $\{I, X\}$ into a logical stochastic Pauli-$Z$ channel.

#### *a. Twirling $|\pi/4\rangle$ magic states*

For target computations that use magic state purification, the output of each purification method is a high-quality $|\pi/4\rangle$ state. These states are then twirled with respect to the gate set $\{I, A\}$, where $A = |\pi/4\rangle\langle\pi/4| - Z|\pi/4\rangle\langle\pi/4|Z$. We now show that this twirling transforms any noise affected the state into a stochastic $Z$ channel.

The logical error rate is defined

$$\epsilon := 1 - \langle\pi/4|\rho|\pi/4\rangle . \tag{B25}$$

Noting that $|\pi/4\rangle\langle\pi/4| - Z|\pi/4\rangle\langle\pi/4|Z = e^{-i\pi/4}SX$ and that

$$e^{-i\pi/4}SX = e^{-i\pi/4}\begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix} = \begin{pmatrix} 0 & e^{-i\pi/4} \\ e^{i\pi/4} & 0 \end{pmatrix} . \tag{B26}$$

Let $|w\rangle := Z|\pi/4\rangle$, the twirled state is then

$$\begin{align}
\frac{1}{2}(\rho + A\rho A^\dagger) = \frac{1}{2}\big((\alpha_1|\pi/4\rangle\langle\pi/4| + \alpha_2|\pi/4\rangle\langle w| + \alpha_3|w\rangle\langle\pi/4| + \alpha_4|w\rangle\langle w|) + A(\alpha_1|\pi/4\rangle\langle\pi/4| + \alpha_2|\pi/4\rangle\langle w| \tag{B27}\\
+ \alpha_3|w\rangle\langle\pi/4| + \alpha_4|w\rangle\langle w|)A^\dagger\big) \tag{B28}\\
= \frac{1}{2}\big(\alpha_1(|\pi/4\rangle\langle\pi/4| + A|\pi/4\rangle\langle\pi/4|A^\dagger) + \alpha_4(|w\rangle\langle w| + A|w\rangle\langle w|A^\dagger)\big) \tag{B29}\\
= \frac{1}{2}\big(\alpha_1(|\pi/4\rangle\langle\pi/4| + |\pi/4\rangle\langle\pi/4|) + \alpha_4(|w\rangle\langle w| + |w\rangle\langle w|)\big) \tag{B30}\\
= \alpha_1|\pi/4\rangle\langle\pi/4| + \alpha_4|w\rangle\langle w| \tag{B31}\\
= (1-\epsilon)|\pi/4\rangle\langle\pi/4| + \epsilon|w\rangle\langle w| . \tag{B32}\\
\tag{B33}
\end{align}$$

Where we have decomposed $\rho$ in the basis of the orthogonal states $|\pi/4\rangle$ and $|w\rangle$ with $\alpha_i \in \mathbb{C}$ and $\sum_i |\alpha_i|^2 = 1$, and then used that

$$\begin{aligned}
&A|\pi/4\rangle\langle w|A^\dagger + |\pi/4\rangle\langle w| \\
&\qquad = (|\pi/4\rangle\langle\pi/4| - |w\rangle\langle w|)|\pi/4\rangle\langle w|(|\pi/4\rangle\langle\pi/4| \\
&\qquad\qquad - |w\rangle\langle w|)^\dagger + |\pi/4\rangle\langle w| \\
&\qquad = -|\pi/4\rangle\langle\pi/4|\pi/4\rangle\langle w|w\rangle\langle w| + |\pi/4\rangle\langle w| \\
&\qquad = 0
\end{aligned} \tag{B34}$$

to get the second equality. The twirled noisy state may then be written as

$$\begin{aligned}
\mathcal{T}(\rho) &= (1-\epsilon)|\pi/4\rangle\langle\pi/4| + \epsilon|w\rangle\langle w| \\
&= (1-\epsilon)|\pi/4\rangle\langle\pi/4| + \epsilon Z|\pi/4\rangle\langle\pi/4|Z.
\end{aligned} \tag{B35}$$

And so the twirled noise is a stochastic $Z$ channel.

#### *b. Twirling $|\pi/2\rangle$ magic states*

We now show that this twirling $|\pi/2\rangle$ magic states transforms any noise affecting these states into stochastic $Z$ channels. These states are twirled with respect to the gate set $\{I, B\}$, where $B = ZX$. The logical error rate is defined

$$\epsilon := 1 - \langle\pi/2|\rho|\pi/2\rangle . \tag{B36}$$

Let $w := Z|\pi/2\rangle$, the twirled state is then

$$\frac{1}{2}(\rho + B\rho B^\dagger) = \frac{1}{2}\big((\alpha_1 |\pi/2\rangle \langle \pi/2| + \alpha_2 |\pi/2\rangle \langle w| + \alpha_3 |w\rangle \langle \pi/2| + \alpha_4 |w\rangle \langle w|) + B(\alpha_1 |\pi/2\rangle \langle \pi/2| + \alpha_2 |\pi/2\rangle \langle w| \tag{B37}$$

$$+ \alpha_3 |w\rangle \langle \pi/2|) + \alpha_4 |w\rangle \langle w|)B^\dagger\big) \tag{B38}$$

$$= \frac{1}{2}\big(\alpha_1(|\pi/2\rangle \langle \pi/2| + B|\pi/2\rangle \langle \pi/2| B^\dagger) + \alpha_4(|w\rangle \langle w| + B|w\rangle \langle w| B^\dagger)\big) \tag{B39}$$

$$= \frac{1}{2}\big(\alpha_1(|\pi/2\rangle \langle \pi/2| + |\pi/2\rangle \langle \pi/2|) + \alpha_4(|w\rangle \langle w| + |w\rangle \langle w|)\big) \tag{B40}$$

$$= \alpha_1 |\pi/2\rangle \langle \pi/2| + \alpha_4 |w\rangle \langle w| \tag{B41}$$

$$= (1-\epsilon)|\pi/2\rangle \langle \pi/2| + \epsilon |w\rangle \langle w| \,. \tag{B42}$$

$$\tag{B43}$$

Similarly to before, $\rho$ is decomposed in the basis of the orthogonal states $|\pi/2\rangle$ and $|w\rangle$ with $\alpha_i \in \mathbb{C}$ and $\sum_i |\alpha_i|^2 = 1$, and then the relation

$$\begin{aligned} &B|\pi/2\rangle \langle w| B^\dagger + |\pi/2\rangle \langle w| \\ &\quad = (|\pi/2\rangle \langle \pi/2| \ - |w\rangle \langle w|)\,|\pi/2\rangle \langle w|\,(|\pi/2\rangle \langle \pi/2| \\ &\qquad - |w\rangle \langle w|)^\dagger + |\pi/2\rangle \langle w| \\ &\quad = -|\pi/2\rangle \langle \pi/2|\pi/2\rangle \langle w|w\rangle \langle w| + |\pi/2\rangle \langle w| \\ &\quad = 0 \end{aligned} \tag{B44}$$

is applied to get the third equality. And so the twirled noise is a stochastic $Z$ channel with the noisy state expressed as

$$\begin{aligned} \mathcal{T}(\rho) &= (1-\epsilon)|\pi/2\rangle \langle \pi/2| + \epsilon |w\rangle \langle w| \\ &= (1-\epsilon)|\pi/2\rangle \langle \pi/2| + \epsilon Z|\pi/2\rangle \langle \pi/2| Z. \end{aligned} \tag{B45}$$

### c. *Twirling $|\pi\rangle$ magic states*

We now show that twirling $|\pi\rangle$ magic states transforms any noise affecting the state into a stochastic $Z$ channel. The logical error rate is defined

$$\epsilon := 1 - \langle \pi| \rho |\pi\rangle \,. \tag{B46}$$

Let $w := Z|\pi\rangle$, the twirled state is then

$$\frac{1}{2}(\rho + X\rho X^\dagger) = \frac{1}{2}\big((\alpha_1 |\pi\rangle \langle \pi| + \alpha_2 |\pi\rangle \langle w| + \alpha_3 |w\rangle \langle \pi| + \alpha_4 |w\rangle \langle w|) + X(\alpha_1 |\pi\rangle \langle \pi| + \alpha_2 |\pi\rangle \langle w| \tag{B47}$$

$$+ \alpha_3 |w\rangle \langle \pi|) + \alpha_4 |w\rangle \langle w|)X^\dagger\big) \tag{B48}$$

$$= \frac{1}{2}\big(\alpha_1(|\pi\rangle \langle \pi| + X|\pi\rangle \langle \pi| X^\dagger) + \alpha_4(|w\rangle \langle w| + X|w\rangle \langle w| X^\dagger)\big) \tag{B49}$$

$$= \frac{1}{2}\big(\alpha_1(|\pi\rangle \langle \pi| + |\pi\rangle \langle \pi|) + \alpha_4(|w\rangle \langle w| + |w\rangle \langle w|)\big) \tag{B50}$$

$$= \alpha_1 |\pi\rangle \langle \pi| + \alpha_4 |w\rangle \langle w| \tag{B51}$$

$$= (1-\epsilon)|\pi\rangle \langle \pi| + \epsilon |w\rangle \langle w| \,. \tag{B52}$$

$$\tag{B53}$$

Similarly to before, $\rho$ is decomposed in the basis of the orthogonal states $|\pi\rangle$ and $|w\rangle$ with $\alpha_i \in \mathbb{C}$ and $\sum_i |\alpha_i|^2 = 1$, and then the relation

$$\begin{aligned} &X|\pi\rangle \langle w| X^\dagger + |\pi\rangle \langle w| \\ &\quad = (|\pi\rangle \langle \pi| \ - |w\rangle \langle w|)\,|\pi\rangle \langle w|\,(|\pi\rangle \langle \pi| \\ &\qquad - |w\rangle \langle w|)^\dagger + |\pi\rangle \langle w| \\ &\quad = -|\pi\rangle \langle \pi|\pi\rangle \langle w|w\rangle \langle w| + |\pi\rangle \langle w| \\ &\quad = 0 \end{aligned} \tag{B54}$$

is applied to get the third equality. And so the twirled noise is a stochastic $Z$ channel with the noisy state expressed as

$$\begin{aligned}\mathcal{T}(\rho) &= (1-\epsilon)\left|\pi\right\rangle\left\langle\pi\right| + \epsilon\left|w\right\rangle\left\langle w\right| \\ &= (1-\epsilon)\left|\pi\right\rangle\left\langle\pi\right| + \epsilon Z\left|\pi\right\rangle\left\langle\pi\right| Z.\end{aligned} \tag{B55}$$

### *Twirling scheme for computations involving $|\theta\rangle$ magic states*

If $|\theta\rangle$ magic states are used in the target circuit, the state preparation noise is transformed into logical stochastic Pauli noise using a Pauli twirling approach. The $|\theta\rangle$ magic states are necessarily unpurified, so this approach does not generalise to purified magic states. In this case, the gate applying the Pauli-$Z$ rotation is Pauli twirled, with the rotation angle reversed if the $X$ or $Y$ twirling operations are applied. The angle reversal is then combined with the angle reversal from the gate twirling described in Appendix B 1 c.

Let the initial ancillary qubit logical state be denoted $\rho$. In the absence of noise, this initial state is $\rho = \left|+\right\rangle\left\langle+\right|$. The state-preparation noise of this state is twirled using the gate set $\{I, X\}$, which transforms the logical state preparation noise into a Pauli-$Z$ channel.

To prepare the $|\theta\rangle$ state, a logical $Z(\theta)$ rotation must be applied to the ancillary qubit. Let the application of the noisy logical $Z$ rotation gate be modelled as $\mathcal{E}\circ Z(\theta)\circ(.)$, where $\mathcal{E}(.)$ is the logical noise channel associated with application of the logical $Z$ rotation gate. Applying the noisy Pauli rotation gate to the state $\rho$ results in the state

$$\rho' = \mathcal{E}\circ Z(\theta)\circ(\rho). \tag{B56}$$

If Pauli twirling operations are included, sandwiching the noisy $Z(\theta)$ gate, instead the state is

$$\begin{aligned}\rho_{\mathcal{T}} &= 4^{-1}\Bigg(\sum_{P_i\in\{I,X,Y,Z\}} P_i\circ\mathcal{E}\circ Z(\theta)\circ P_i\Bigg)\circ(\rho) \\ &= 4^{-1}\Bigg(\sum_{P_i\in\{I,X,Y,Z\}} P_i\circ\mathcal{E}\circ P_i\Bigg) \\ &\qquad \circ Z\big((-1)^{b_{P_i}\odot b_Z}\cdot\theta\big)\circ(\rho),\end{aligned} \tag{B57}$$

where $b_{P_i}$ and $b_Z$ are the binary symplectic representations of the Pauli operators $P_i$ and $Z$ respectively. To get the second equality, the Pauli operators have been propagated through the $Z(\theta)$ gate to isolate the logical error channel for the Pauli twirling. However, the Pauli operators that do not commute with $Z$ reverse the rotation angle of the $Z$ rotation gate. This change in rotation angle can be counteracted by updating the rotation angle in each twirling instance. Since the rotation angle depends on the action of physical rotation gates, and we are assuming physical single-qubit gate noise is gate-independent, the updates can be made without changing the gate noise. The twirling operation then becomes

$$\begin{aligned}\rho_{\mathcal{T}} &= 4^{-1}\Bigg(\sum_{P_i\in\{I,X,Y,Z\}} P_i\circ\mathcal{E}\circ Z\big((-1)^{b_{P_i}\odot b_Z}\cdot\theta\big)\circ P_i\Bigg) \\ &\qquad\circ(\rho) \\ &= 4^{-1}\Bigg(\sum_{P_i\in\{I,X,Y,Z\}} P_i\circ\mathcal{E}\circ P_i\Bigg)\circ Z(\theta)\circ(\rho).\end{aligned} \tag{B58}$$

The logical rotation gate channel has now been isolated from the gate, and the transformation to logical stochastic Pauli noise follows in the same way as in Appendix B 1 c. The Pauli twirl eliminates all of the off-diagonal components of the Pauli decomposition of the channel, leaving only the diagonal elements. Taking the Pauli twirl of the channel $\mathcal{E}$ transforms it to

$$\mathcal{E}'(\rho) = \sum_{P_i\in\{I,X,Y,Z\}} c_{P_i} P_i\circ Z(\theta)(\rho). \tag{B59}$$

Where we have defined the set of Pauli operator probability coefficients $\{c_{P_i}\}_i$, such that the coefficient for Pauli operator $P_i$ is $c_{P_i}$. Therefore, the magic state preparation noise is transformed by the twirling into logical stochastic Pauli noise.

In order to combine this with the gate twirling described in Appendix B 1 c, the rotation gate applied in each instance is actually of the form: $Z((-1)^{a+a'}\cdot\theta)$. Where $a$ is the factor from the magic state twirling, with $a = b_{P_i}\odot b_Z$, and $a'$ is the factor from the gate twirling in Appendix B 1 c, with $a' = b_{P_i'}\odot b_P$ where $P_i'$ is the specific instance of the gate twirling used and $P$ is the Pauli rotation basis of the gate.

The corresponding trap magic states are twirled using the same approach.

## Appendix C: Computing the upper-bound on the TVD of the experimental target circuit output from the ideal output

We now derive Theorem 1 from the main text.

To implement the certification protocol, the target and trap circuits are run using encoded logical qubits on the specified quantum device, and the measured outputs of the trap circuits are used to certify the target computation in the following way. The logical twirling scheme described in Section B 1 transforms general logical noise into stochastic Pauli noise, and so the state generated by running one of the trap circuits or the target circuit may be decomposed as

$$\rho^{(i)}_{out} = (1-p_{err})\rho^{(i)}_{out,id} + p^{(i)}_{err}\rho^{(i)}_{noisy}, \tag{C1}$$

where $i\in\{1,...,M+1\}$ is an index indicating which circuit has been run, $p^{(i)}_{err}$ is the probability of at least one logical error occurring at any point in the circuit,

$\rho^{(i)}_{out,id}$ is the ideal output state resulting from running circuit $i$ in the absence of noise, and $\rho^{(i)}_{noisy}$ is the output state of circuit $i$ encompassing the effects of circuit noise. The gate operations of the ideal $i$-th logical circuit, $\mathcal{C}^{(i)}$, may be written as a sequence of $D$ logical gate layers

$$\mathcal{C}^{(i)} = \prod_{j=1}^{D} \mathcal{L}_j^{(i)}, \tag{C2}$$

where $\mathcal{L}_j^{(i)}$ denotes the $j$-th logical gate layer of the $i$-th circuit. Now, including logical circuit noise, the previous term becomes

$$\tilde{\mathcal{C}}^{(i)} = \mathcal{E}^{(i)}_{D+1} \cdot \Big( \prod_{j=1}^{D} \mathcal{E}_j^{(i)} \mathcal{L}_j^{(i)} \Big) \cdot \mathcal{E}_0^{(i)}, \tag{C3}$$

where $\mathcal{E}_j^{(i)}$ denotes the logical effective Pauli noise channel associated with logical gate layer $\mathcal{L}_j^{(i)}$. The noise channels $\mathcal{E}_0^{(i)}$ and $\mathcal{E}_{D+1}^{(i)}$ denote logical state preparation and measurement noise, respectively. Noise arising from the use of noisy magic states and from incorrect decoding during cycles of error correction are included in the set of gate layer noise channels $\{\mathcal{E}_j^{(i)}\}_j$. The total logical error rate of $\tilde{\mathcal{C}}^{(i)}$ is then

$$p^{(i)}_{err} = \sum_{j=0}^{D+1} \sum_{P \in \{I,X,Y,Z\}^{\otimes n}/I^{\otimes n}} p^{(i)}_{P,j}, \tag{C4}$$

where $\{p^{(i)}_{P,j}\}_{P \in \{I,X,Y,Z\}^{\otimes n}/I^{\otimes n}}$ is the set of Pauli coefficients for non-trivial operators of the $j$-th gate layer noise channel for the $i$-th logical circuit.

After the randomly ordered target and trap circuits are run on the quantum device and the measured outputs are recorded, the outputs of the trap circuits are classically postprocessed to compute an upper-bound on the error of the target circuit output. Specifically, a bound on the TVD between the experimentally sampled target circuit output distribution, $\mathcal{D}_{exp}$, and the ideal distribution, $\mathcal{D}_{ideal}$. If the states $\rho_{out}$ and $\rho_{out,id}$ are measured in the computational basis, for which the measurement projection operators are $\{\Pi_s\}_s$ where $\Pi_s = |s\rangle\langle s|$ for $s \in \{0,1\}^n$, the TVD of the output distributions can be expressed as

$$\begin{aligned} \delta(\mathcal{D}_{exp}, \mathcal{D}_{ideal}) &= \frac{1}{2} \sum_{s \in \{0,1\}^n} \big| \mathrm{Tr}\big[(\rho_{out} - \rho_{out,id})\Pi_s\big] \big| \\ &= \frac{1}{2} \sum_{s \in \{0,1\}^n} |p_{exp}(s) - p_{ideal}(s)|. \end{aligned} \tag{C5}$$

Now let the target circuit be denoted by index 1, and the indices $i \in \{2, \ldots, M+1\}$ correspond to trap circuits. Let the probability that the $i$-th trap circuit returns a bit string other than $0^n$ be denoted by $p^{(i)}_{inc}$, the probability that errors occur in a trap circuit and cancel be denoted by $p_{canc}$, and the probability that if a single error affects a trap circuit it is detected be denoted by $p_{det}$. The probability that a trap circuit does not output the bit string $0^n$, i.e. that errors occur and are detected, is lower-bounded by the probability that errors occur in a trap circuit, do not cancel, and are detected, so that

$$p^{(i)}_{inc} \geq p^{(i)}_{err} \cdot (1 - p_{canc}) \cdot p_{det}, \tag{C6}$$

which may be rearranged to

$$p^{(i)}_{err} \leq \frac{p^{(i)}_{inc}}{(1 - p_{canc}) \cdot p_{det}}. \tag{C7}$$

The TVD can be upper-bounded using the inequality: $\delta(\mathcal{D}^{(i)}_{exp}, \mathcal{D}^{(i)}_{ideal}) \leq D(\rho^{(i)}_{out}, \rho^{(i),id}_{out})$, where $D(.,.)$ is the trace distance. The trace distance between the experimental and ideal output states of the $i$-th circuit is

$$\begin{aligned} D(\rho^{(i)}_{out}, \rho^{(i)}_{out,id}) &= \frac{1}{2}\mathrm{Tr}(|\rho^{(i)}_{out} - \rho^{(i)}_{out,id}|_1) \\ &= \frac{1}{2}\mathrm{Tr}(|(1 - p^{(i)}_{err})\rho^{(i)}_{out,id} \\ &\qquad + p^{(i)}_{err}\rho^{(i)}_{noisy} - \rho^{(i)}_{out,id}|_1) \\ &= \frac{1}{2} p^{(i)}_{err} \mathrm{Tr}(|\rho^{(i)}_{noisy} - \rho^{(i)}_{out,id}|_1) \end{aligned} \tag{C8}$$

And because $\frac{1}{2}\mathrm{Tr}(|\rho^{(i)}_{noisy} - \rho^{(i)}_{out,id}|_1) \leq 1$, it follows that $D(\rho^{(i)}_{out}, \rho^{(i)}_{out,id}) \leq p^{(i)}_{err}$. Therefore, the TVD bounded of the $i$-th circuit is bounded by

$$\delta(\mathcal{D}^{(i)}_{exp}, \mathcal{D}^{(i)}_{ideal}) \leq \frac{p^{(i)}_{inc}}{(1 - p_{canc}) \cdot p_{det}}. \tag{C9}$$

The noise channels affecting different circuit runs can be different, and are assumed to be independent, as stated in assumption A3. The output of the trap circuits may be used to upper-bound the average total circuit error rate over the ensemble of possible noise behaviours affecting the target and trap circuits. The target circuit can be viewed as subject to the distribution of all noise behaviours, with an overall error rate equal to the expected error rate of the distribution. Therefore, the upper-bound on the total circuit error rate computed from the trap circuits may be applied to the target circuit. That is

$$\delta(\mathcal{D}^{(0)}_{exp}, \mathcal{D}^{(0)}_{ideal}) \leq \frac{\sum_{i>1} p^{(i)}_{inc}}{M \cdot (1 - p_{canc}) \cdot p_{det}}. \tag{C10}$$

The value of $M^{-1}\sum_{i>1} p^{(i)}_{inc}$ is experimentally estimated by $N_{inc}/M$, where $N_{inc}$ in the number of experimental trap circuit runs for which an error is detected. The independence of the circuit runs means that the output of each trap circuit can be modelled as an independent Bernoulli random variable, where the outcome 0 indicates that the $0^n$ string is measured at the output, and the outcome 1 indicates that any other bit string is measured. Let $X_1, X_2, \ldots, X_M$ with $X_i \sim \mathrm{Bern}(p_i)$ be independent Bernoulli random variables modelling the $M$ trap circuit outputs, where $X_i \in \{0,1\}$ for all $i$, and $\mathbb{E}(X_i) = p^{(i)}_{inc}$. The $p^{(i)}_{inc}$ values are drawn from a distribution $\mathcal{D}_{p^{(i)}_{inc}}$ representing all possible noise behaviours that can affect the circuits, and the expectation of this distribution is denoted by $\mathbb{E}(p^{(i)}_{inc})$. The empirical mean of the $M$ random variables is then $\bar{X}_M = M^{-1}\sum_{i>1} X_i = N_{inc}/M$ and the expected mean is $\mathbb{E}(\bar{X}_M) = M^{-1}\sum_{i>1} p^{(i)}_{inc}$. Now

$$|\bar{X}_M - \mathbb{E}(p^{(i)}_{inc})| \leq |\bar{X}_M - \mathbb{E}(\bar{X}_M)| + |\mathbb{E}(\bar{X}_M) - \mathbb{E}(p^{(i)}_{inc})| \tag{C11}$$

by a triangle inequality. The two expressions on the RHS can now be upper-bounded. By Hoeffding's inequality

$$\Pr\big(|\bar{X}_M - \mathbb{E}(\bar{X}_M)| \leq \epsilon\big) \geq 1 - 2e^{-2\epsilon^2 M}. \tag{C12}$$

And, also by Hoeffding's inequality,

$$\Pr\big(|\mathbb{E}(\bar{X}_M) - \mathbb{E}(p^{(i)}_{inc})| \leq \epsilon\big) \geq 1 - 2e^{-2\epsilon^2 M}. \tag{C13}$$

These can be combined using a union bound, to get

$$\Pr\big(|\bar{X}_M - \mathbb{E}(p^{(i)}_{inc})| \leq \epsilon\big) \geq 1 - 4e^{-\frac{\epsilon^2 M}{2}}. \tag{C14}$$

To achieve at least $\alpha$-confidence requires $M > \frac{2}{\epsilon^2}\log\Big(\frac{4}{1-\alpha}\Big)$ trap circuits. This is obtained by solving the inequality $\alpha \leq 1 - 4e^{-\frac{\epsilon^2 M}{2}}$ for $M$.

The probability that a single error affects a trap circuit and is detected, $p_{det}$, is lower-bounded due to the following lemma:

**Lemma 3.** *Any single logical Pauli error of arbitrary weight occurring at any single time-step during the computation is detected with probability $p_{det} \geq 1/2$.*

This result is derived in Appendix D.

The probability that multiple errors affect a trap circuit and cancel, $p_{canc}$, may be upper-bounded using one of the following three lemmas derived in Appendix E. In Lemma 4, the probability of error cancellation is neglected. In Lemma 5, the probability of error cancellation from errors affecting multi-qubit Clifford gates and gates performed using magic states is bounded, and any other possible error cancellation is neglected. In Lemma 6, the probability of error cancellation for any collection of errors is bounded. Lemmas 4 and 5 use the original trap construction. While Lemma 6 uses a modified trap circuit construction, this is described in Appendix E 3.

**Lemma 4.** *In the noise regime where $N_{\mathcal{E}} q_{max} \ll 1$, the probability of error cancellation is bounded*

$$p_{canc} \leq O((N_{\mathcal{E}} q_{max})^2), \tag{C15}$$

*where $N_{\mathcal{E}}$ is the number of logical noise channels affecting the circuit, and $q_{max}$ is the highest total channel error rate of the noise channels. In this regime, it may be assumed that $p_{canc} \leq \beta$, where $\beta = 0$. If this $\beta$ value is used to compute the TVD upper-bound, $\gamma$, the magnitude of the approximation error from this assumption is bounded by $|\epsilon_\gamma| \leq O((N_{\mathcal{E}} q_{max})^2)$.*

This result is derived in Appendix E 1.

**Lemma 5.** *In the regime where the dominant source of noise is due to the $\mathcal{J}$-type gate layers, the probability of error cancellation is bounded by*

$$p_{canc} \leq 1/2 + O(N_{\mathcal{J}} \cdot q_{\mathcal{J},max} \cdot N_{\mathcal{C}} \cdot q_{\mathcal{C},max}), \tag{C16}$$

*where $N_{\mathcal{J}}$ is the number of logical noise channels affecting $\mathcal{J}$-type gate layers in the circuit and $q_{\mathcal{J},max}$ is the highest total channel error rate of these, and $N_{\mathcal{C}}$ is the number of logical noise channels affecting logical single-qubit Clifford gate layers, and state preparation and measurement, and $q_{\mathcal{C},max}$ is the highest of these. In this regime, it may be assumed that $p_{canc} \leq \beta$, where $\beta = 1/2$. If this $\beta$ value is used to compute the TVD upper-bound, $\gamma$, the magnitude of the approximation error from this assumption is bounded by*

$$|\epsilon_\gamma| \leq O\Big(N_{\mathcal{J}} \cdot q_{\mathcal{J},max} \cdot N_{\mathcal{C}} \cdot q_{\mathcal{C},max}\Big). \tag{C17}$$

This result is derived in Appendix E 2.

The following lemma uses a slightly different trap circuit construction, referred to as *the modified trap circuit construction*, described in Appendix E 3.

**Lemma 6.** *For the modified trap circuit construction, the probability of error cancellation in the trap circuits is upper bounded by $p_{canc} \leq 7/8$.*

This result is derived in Appendix E 3. It can be shown that the same error detection bound stated in Lemma 3 also applies for the modified trap circuits; i.e. $p_{det} \geq 1/2$.

So that, using Lemma 3, $p_{det} \geq 1/2$, and $p_{canc} \leq \beta$, where

$$\beta = \begin{cases} 0 & \text{if Lemma 4 is used} \\ 1/2 & \text{if Lemma 5 is used} \\ 7/8 & \text{if Lemma 6 is used,} \end{cases} \tag{C18}$$

the TVD is bounded by

$$\delta(\mathcal{D}_{exp}, \mathcal{D}_{ideal}) \leq \frac{2p_{inc}}{1-\beta}. \tag{C19}$$

The soundness of the protocol comes from upper bounding the false positive rate (i.e. $1 - (1 - p_{canc})p_{det}$) with a soundness parameter $\epsilon_s$. The different values of $\beta$ and Lemma 3 provide soundness bounds of

$$\epsilon_s = \begin{cases} 1/2 & \text{if Lemma 4 is used} \\ 3/4 & \text{if Lemma 5 is used} \\ 15/16 & \text{if Lemma 6 is used.} \end{cases} \tag{C20}$$

Let $\gamma = \frac{2p_{inc}}{1-\beta}$, the bound

$$\frac{1}{2}\sum_{s\in\{0,1\}^n} |p_{exp}(s) - p_{ideal}(s)| \leq \gamma, \tag{C21}$$

then applies with a user defined accuracy $\epsilon$ and confidence $\alpha$, where the number of trap circuits satisfies the inequality $M \geq \frac{2}{\epsilon^2}\log\left(\frac{4}{1-\alpha}\right)$. That is, at least $\lceil \frac{2}{\epsilon^2}\log\left(\frac{4}{1-\alpha}\right)\rceil$ traps are needed to ensure a confidence of $\alpha$ that the $\gamma$-estimator is within $\pm\epsilon$ of the ideal value.

This bound may be used to upper-bound the expectation value error of measurement of an observable $O$. Let $\rho_{exp} = \sum_s p_{exp}(s)\,|s\rangle\,\langle s|$ be a mixed state representing the possible experimental measurement outcomes for the target circuit. We have that

$$\begin{aligned}\langle O\rangle_{\rho_{exp}} &= \mathrm{Tr}[\rho_{exp}O] \\ &= \sum_s p_{exp}(s)\,\langle s|\,O\,|s\rangle \\ &= \sum_s p_{exp}(s)\langle O\rangle_s.\end{aligned} \tag{C22}$$

Similarly, for the ideal measurement outcomes, we have that

$$\langle O\rangle_{\rho_{ideal}} = \sum_s p_{ideal}(s)\langle O\rangle_s. \tag{C23}$$

The expectation value error can then be upper-bounded, as

$$\begin{aligned}|\langle O\rangle_{\rho_{exp}} - \langle O\rangle_{\rho_{ideal}}| &= |\sum_{s\in\{0,1\}^n}(p_{exp}(s) - p_{ideal}(s))\langle O\rangle_s| \\ &\leq \sum_{s\in\{0,1\}^n}|(p_{exp}(s) - p_{ideal}(s))|\cdot|\langle O\rangle_s| \\ &\leq 2\delta(\mathcal{D}_{exp}, \mathcal{D}_{ideal})\|\langle O\rangle_s\| \\ &\leq 2\gamma\|O\|,\end{aligned} \tag{C24}$$

where $\|.\|$ is the operator norm, also called the spectral norm, and $||O|| = \max_{|\psi\rangle}|\,\langle\psi|\,O\,|\psi\rangle\,| = \max_i|\lambda_i|$ for any Hermitian operator $O$. The third inequality follows from that $|\,\langle s|\,O\,|s\rangle\,| \leq ||O||$ for all $|s\rangle$, since the operator norm of a Hermitian operator is its largest absolute eignenvalue.

## Appendix D: Lower-bound for the probability that trap circuits detect any single error

We now prove Lemma 3, that states:

*Any single logical Pauli error of arbitrary weight occurring at any single time-step during the computation is detected with probability $p_{det} \geq 1/2$.*

*Proof.* Let the set of all possible trap circuits for a given target circuit be denoted $\{\mathcal{C}_T^{(i)}\}_i$. The $i$-th noisy trap circuit, $\tilde{\mathcal{C}}_T^{(i)}$, may be decomposed into logical gate layers and noise channels, and, including state preparation and measurement noise, expressed in the form

$$\tilde{\mathcal{C}}_T^{(i)} = \mathcal{E}_{D+1}\cdot\prod_{j=1}^{D}(\mathcal{E}_j\cdot\mathcal{L}_j^{(i)})\cdot\mathcal{E}_0, \tag{D1}$$

where $\mathcal{E}_j$ denotes the logical Pauli noise channel associated with logical gate layer $\mathcal{L}_j^{(i)}$, $\mathcal{E}_0$ denotes logical state preparation noise and $\mathcal{E}_{D+1}$ logical measurement noise. Explicitly labelling gate layers where logical Pauli twirling operations are performed, an ideal trap circuit may be written

$$\mathcal{C}_T^{(i)} = \mathcal{P}_{D/3+1}^{(1)}\cdot\prod_{m=1}^{D/3}\left(\mathcal{P}_{3m}^{(2)}\cdot\mathcal{L}_{3m}^{(i)}\cdot\mathcal{P}_{3m}^{(1)}\right)\cdot\mathcal{P}_0^{(2)}. \tag{D2}$$

Each trap circuit is generated by randomly choosing whether to apply a layer of Hadamard gates at the beginning and end of the circuit, and randomly generating the sandwiching gate layers. For the $i$-th trap circuit, let the $j$-th sandwiching gate layer be denoted $\mathcal{W}_j^{(i)} \in \{S, S^\dagger, H\}^{\otimes n}$. Let $\mathcal{H} = H^{\otimes n}$, and let $R_{\mathcal{W},l}$ denote a specific instance of sandwiching gate layers for a trap circuit, and let $\{R_{\mathcal{W},l}\}_l$ denote the set of all possible sandwiching gate layer instances. Let $\mathcal{J}_j$ denote the $j$-th logical gate layer that can contain gates implemented using trap magic states, and also multi-qubit Clifford gates. As $t \in \{0,1\}$ and $R_{\mathcal{W},l} \in \{R_{\mathcal{W},l}\}_l$, there are then $|\{\mathcal{C}_T^{(i)}\}_i| = 2\cdot|\{R_{\mathcal{W},l}\}_l|$ possible trap circuits. The $i$-th trap circuit from the set $\{\mathcal{C}_T^{(i)}\}_i$ may be written

$$\mathcal{C}_T^{(i)} = \mathcal{P}_{K+2}^{(1)}\cdot\mathcal{P}_{K+1}^{(2)}\cdot\mathcal{H}^t\cdot\mathcal{P}_{K+1}^{(1)}\cdot\prod_{k=1}^{K}\left(\mathcal{P}_{3k+1}^{(2)}\cdot\mathcal{W}_{3k+1}^{(i)}\cdot\mathcal{P}_{3k+1}^{(1)}\cdot\mathcal{P}_{3k}^{(2)}\cdot\mathcal{J}_{3k}\cdot\mathcal{P}_{3k}^{(1)}\cdot\mathcal{P}_{3k-1}^{(2)}\cdot\mathcal{W}_{3k-1}^{(i)}\cdot\mathcal{P}_{3k-1}^{(1)}\right) \tag{D3}$$

$$\cdot\,\mathcal{P}_1^{(2)}\cdot\mathcal{H}^t\cdot\mathcal{P}_1^{(1)}\cdot\mathcal{P}_0^{(2)}. \tag{D4}$$

$$\tag{D5}$$

In practice, when each such circuit is run adjacent Pauli layers are compiled together, leading to $D$ circuit layers overall.

We now ignore Pauli twirling operations, as these do not change the circuit logic but rather serve to transform general logical noise into stochastic logical Pauli noise (see Appendix B). However, the noise channels associated with the Pauli gate layers are combined with the noise channel associated with the previous time-step non-Pauli gate layer, and so are included in the following analysis. Removing the Pauli gate layers from the previous expression, it becomes

$$\mathcal{C}_T^{(i)\prime} = \mathcal{H}^t \cdot \prod_{k=1}^{K} (\mathcal{W}_{3k+1}^{(i)} \cdot \mathcal{J}_{3k}^{(i)} \cdot \mathcal{W}_{3k-1}^{(i)}) \cdot \mathcal{H}^t. \tag{D6}$$

The different choices of randomising gates define the different trap circuits. As $t \in \{0,1\}$ and $R_{\mathcal{W},l} \in \{R_{\mathcal{W},l}\}_l$ are both chosen with uniform probability over their respective sets, the probability of measuring $s = 0^n$ in the absence of logical noise, summing over all possible trap circuits, is

$$\begin{aligned}\Pr(s = 0^n) &= \frac{1}{2 \cdot |R_{\mathcal{W},l}\}_l|} \sum_{t \in \{0,1\}} \sum_{R_{\mathcal{W},l} \in \{R_{\mathcal{W},l}\}_l} \langle 0^n | \, \mathcal{H}^t \bigcirc_{k=1}^{K} (\mathcal{W}_{3k+1}^{(i)} \circ \mathcal{J}_{3k} \circ \mathcal{W}_{3k-1}^{(i)}) \circ \mathcal{H}^t \circ (|0^n\rangle \langle 0^n|) \, |0^n\rangle & \text{(D7)} \\ &= 1. & \text{(D8)}\end{aligned}$$

Where the final line follows from the fact that the $\mathcal{J}_j$ layers consist of operations that stabilize the input state. As they can include identity operations (from the use of trap magic states instead of target magic states when performing magic state-requiring gates), and C$Z$ and CNOT gates. This means that $\mathcal{W}_{3k+1}\mathcal{J}_{3k}^{(i)}\mathcal{W}_{3k-1}\mathcal{H}^t |0^n\rangle = \mathcal{H}^t |0^n\rangle$, for all $k \in \{1, \ldots, K\}$ and $t \in \{0,1\}$. The result is then that no error is registered as having occurred during any ideal trap computation with probability 1.

We now consider noise channels at different time-steps in the computation leading to errors, lower bounding the detection probability over the set of possible trap circuits. Including a logical computational state preparation Pauli noise channel, $\mathcal{E}_0$, the probability of the trap circuits not detecting an error is then

$$\Pr(s = 0^n | \mathcal{E}_0) = \frac{1}{2 \cdot |R_{\mathcal{W},l}\}_l|} \sum_{t \in \{0,1\}} \sum_{R_{\mathcal{W},l} \in \{R_{\mathcal{W},l}\}_l} \langle 0^n | \, \mathcal{H}^t \bigcirc_{k=1}^{K} (\mathcal{W}_{3k+1}^{(i)} \circ \mathcal{J}_{3k}^{(i)} \circ \mathcal{W}_{3k-1}^{(i)}) \circ \mathcal{H}^t \circ \mathcal{E}_0 \circ (|0^n\rangle \langle 0^n|) \, |0^n\rangle \, . \tag{D9}$$

$$\tag{D10}$$

Of the possible Pauli errors that can affect the trap circuit due to the noise channel $\mathcal{E}_0$, only those that have a non-trivial $X$ or $Y$ component will not stabilise the quantum state. When an error of this type occurs then $p(s = 0^n | P_0) = \langle 0^n | \, P_0 \, |0^n\rangle = 0$, where $P_0$ is a Pauli error with a non-trivial $X$ or $Y$ component. Meaning any Pauli error that non-trivially changes the prepared computational state is detected with probability 1. Similar arguments can be made regarding logical computational measurement Pauli noise, as errors changing the measured output may also be detected with probability 1.

Now, considering logical noise affecting the first Hadamard gate layer, the probability of trap circuits not detecting an error is

$$\begin{aligned}\Pr(s = 0^n | \mathcal{E}_1) &= \frac{1}{2 \cdot |R_{\mathcal{W},l}\}_l|} \sum_{t \in \{0,1\}} \sum_{R_{\mathcal{W},l} \in \{R_{\mathcal{W},l}\}_l} \langle 0^n | \, \mathcal{H}^t \bigcirc_{k=1}^{L} (\mathcal{W}_{3k+1}^{(i)} \circ \mathcal{J}_{3k}^{(i)} \circ \mathcal{W}_{3k-1}^{(i)}) \circ \mathcal{E}_1 \circ \mathcal{H}^t \circ (|0^n\rangle \langle 0^n|) \, |0^n\rangle & \\ & & \text{(D11)} \\ &= \frac{1}{2 \cdot |R_{\mathcal{W},l}\}_l|} \sum_{t \in \{0,1\}} \sum_{R_{\mathcal{W},l} \in \{R_{\mathcal{W},l}\}_l} \langle 0^n | \, \mathcal{H}^t \circ \mathcal{E}_1 \circ \mathcal{H}^t \circ (|0^n\rangle \langle 0^n|) \, |0^n\rangle & \text{(D12)} \\ &< \frac{1}{2}. & \text{(D13)}\end{aligned}$$

Since $t$ is 0 or 1, each with probability 1/2, if a Pauli error $P_1$ occurs with a non-trivial $Y$ component, it is detected with probability 1. If $P_1$ has a non-trivial $X$ component, but a trivial $Y$ and $Z$ component, it is detected with probability 1/2. Likewise, if $P_1$ has a non-trivial $Z$ component, but a trivial $X$ and $Y$ component, it is detected with probability 1/2. Meaning any Pauli error is detected with probability at least 1/2. Similar arguments can be applied to Pauli noise affecting the final Hadamard layer.

Now considering noise affecting one of the $\mathcal{W}_j^{(i)}$ gate layers, the probability that noise affects gate layer $\mathcal{W}_{3M-1}^{(i)}$ and is not detected is

$$
\begin{aligned}
\Pr(s = 0^n|\mathcal{E}_{3M-1}) &= \frac{1}{2\cdot|R_{\mathcal{W},l}\}_l|}\sum_{t\in\{0,1\}}\sum_{R_{\mathcal{W},l}\in\{R_{\mathcal{W},l}\}_l} \langle 0^n|\,\mathcal{H}^t \bigcirc_{k=M+1}^{L} (\mathcal{W}_{3k+1}^{(i)}\circ\mathcal{J}_{3k}^{(i)}\circ\mathcal{W}_{3k-1}^{(i)}) && \text{(D14)}\\
&\quad \circ(\mathcal{W}_{3M+1}^{(i)}\circ\mathcal{J}_{3M}^{(i)}\circ\mathcal{E}_{3M-1}\circ\mathcal{W}_{3M-1}^{(i)}) \bigcirc_{k=1}^{M-1} (\mathcal{W}_{3k+1}^{(i)}\circ\mathcal{J}_{3k}^{(i)}\circ\mathcal{W}_{3k-1}^{(i)}) && \text{(D15)}\\
&\quad \circ\mathcal{H}^t\circ(|0^n\rangle\langle 0^n|)\,|0^n\rangle && \text{(D16)}\\
&= \frac{1}{2\cdot|R_{\mathcal{W},l}\}_l|}\sum_{t\in\{0,1\}}\sum_{R_{\mathcal{W},l}\in\{R_{\mathcal{W},l}\}_l} \langle 0^n|\,\mathcal{H}^t \bigcirc_{k=M}^{L} (\mathcal{W}_{3k+1}^{(i)}\circ\mathcal{J}_{3k}^{(i)}\circ\mathcal{W}_{3k-1}^{(i)}) && \text{(D17)}\\
&\quad \circ\mathcal{E}_{3M-1}{}' \bigcirc_{k=1}^{M-1} (\mathcal{W}_{3k+1}^{(i)}\circ\mathcal{J}_{3k}^{(i)}\circ\mathcal{W}_{3k-1}^{(i)})\circ\mathcal{H}^t\circ(|0^n\rangle\langle 0^n|)\,|0^n\rangle && \text{(D18)}\\
&= \frac{1}{2\cdot|R_{\mathcal{W},l}\}_l|}\sum_{t\in\{0,1\}}\sum_{R_{\mathcal{W},l}\in\{R_{\mathcal{W},l}\}_l} \langle 0^n|\,\mathcal{H}^t\circ\mathcal{E}_{3M-1}{}'\circ\mathcal{H}^t\circ(|0^n\rangle\langle 0^n|)\,|0^n\rangle && \text{(D19)}\\
&< \frac{1}{2}. && \text{(D20)}
\end{aligned}
$$

Where $\mathcal{E}_{3M-1}\mathcal{W}_{3M-1}^{(i)} = \mathcal{W}_{3M-1}^{(i)}\mathcal{E}_{3M-1}{}'$, with $\mathcal{E}_{3M-1}{}'$ being another Pauli channel, since $\mathcal{W}_{3M-1}^{(i)}\in\{S, S^\dagger, H\}^{\otimes n}$. Using similar arguments as were applied for the case of errors affecting the Hadamard gate layers, any error occurring as a result of this error channel is detected with probability at least 1/2. Similar arguments can be made about errors affecting one of the $\mathcal{J}_j^{(i)}$ layers. Therefore, any single logical Pauli error occurring at any single timestep during a trap computation is detected with probability $p_{det}\geq 1/2$. □

## Appendix E: Upper-bounds for the probability of error cancellation in trap circuits

We now wish to bound the probability that errors affect a trap circuit and cancel, rendering them undetectable. Combined with the bound derived in the previous section on the trap circuit detection probability of individual errors, this then provides an overall bound on the detection probability of any collection of errors affecting a trap circuit; see eqn. C19. We consider three different scenarios, deriving bounds on the probability of error cancellation in the trap circuits, $p_{canc}$, for each. The three scenarios are:

1. The probability of logical error cancellation is sufficiently small that cancellation effects can be reasonably neglected.
2. The logical error rates for performing gates using magic states are appreciably larger than those for performing Clifford gates, state preparation or measurement. The leading order cancellation effects are then from errors affecting gates performed using magic states, while other cancellation effects can be reasonably neglected.
3. No cancellation effects are neglected.

Scenario (1) is appropriate when it is assumed that the noise channels affecting circuits are Markovian, as there would be no time-correlated noise effects. Scenario (2) is appropriate when it is assumed that the noise from $\mathcal{J}$-type gate layers, i.e. the noise from multi-qubit Clifford gates and gates requiring magic states, is non-Markovian, while noise from single-qubit Clifford gates is Markovian. In Scenario (3), no assumptions need to be made about the Markovianity of the noise. So Scenario (1) considers complete Markovianity, Scenario (2) part Markovianity and part non-Markovianity, and Scenario (3) potentially complete non-Markovianity. This is because Pauli twirled non-Markovian noise results in classically correlated stochastic Pauli channels [56, 57]. And this can lead to time-correlated Pauli errors affecting the trap circuits and potentially cancelling. In Scenario (2), it is shown that the randomisation of the trap circuits means that the probability of errors affecting $\mathcal{J}$-type layers and cancelling can be upper-bounded. In Scenario (3), it is shown that, using a modified version of the trap circuit construction, the probability of any errors affecting a trap circuit and cancelling can be upper-bounded.

In the Section E 1 analysis, Scenario (1) provides the smallest upper-bound value on the probability of error cancellation, but the assumption is required that errors do not cancel in the trap circuits. The next smallest upper-bound value is derived in Section E 2 for Scenario (2), here the probability of error cancellation due to errors affecting gates performed using magic states is bounded, and the assumption is required that there are no cancellation effects due to errors from other components of the trap circuit. No assumptions are made about error cancellation for the Scenario (3) analysis presented in Section E 3; however, the price for this is a larger upper-bound. The logical randomised compiling described in Appendix B transforms general logical noise to logical stochastic Pauli noise. This means that, in the following error cancellation analysis, the Pauli channels affecting a trap circuit may be propagated together to form a single Pauli channel. And a bound on the error cancellation probability may be derived by analysing the components of the error channels that cancel when they are combined.

### 1. Neglecting the error cancellation

If the total error rates of each logical noise channel are orders of magnitude smaller than one, then the likelihood of error cancellation becomes negligibly small and can reasonably be ignored. The expected number of errors when running a circuit, $N_e$, is bounded $N_e \leq N_{\mathcal{E}} q_{max}$, where $N_{\mathcal{E}}$ is the number of logical noise channels affecting the circuit, and $q_{max}$ is the highest total channel error rate of these noise channels. Only in the regime where $N_{\mathcal{E}} q_{max} \ll 1$ is it possible to achieve a non-trivial computational output. When performing logical computation in this logical noise regime, it is reasonable to assume that error cancellation happens with sufficiently low probability that it can be neglected when computing the TVD bound. This is formalised in the following lemma.

We now prove Lemma 4, that states:

*In the noise regime where $N_{\mathcal{E}} q_{max} \ll 1$, the probability of error cancellation is bounded*

$$p_{canc} \leq O((N_{\mathcal{E}} q_{max})^2), \tag{E1}$$

*where $N_{\mathcal{E}}$ is the number of logical noise channels affecting the circuit, and $q_{max}$ is the highest total channel error rate of the noise channels. In this regime, it may be assumed that $p_{canc} \leq \beta$, where $\beta = 0$. If this $\beta$ value is used to compute the TVD upper-bound, $\gamma$, the magnitude of the approximation error from this assumption is bounded $|\epsilon_\gamma| \leq O((N_{\mathcal{E}} q_{max})^2)$.*

*Proof.* If there are $N_{\mathcal{E}}$ stochastic Pauli channels that affect each trap circuit, the set of total error rates for these channels is denoted by $\{q_{tot}(j)\}_{j=1}^{N_{\mathcal{E}}}$, and the highest error rate is $q_{max} = \max(\{q_{tot}(j)\}_{j=1}^{N_{\mathcal{E}}})$. The probability of cancellation is upper bounded

$$p_{canc} \leq \sum_{j=2}^{N_{\mathcal{E}}} \binom{N_{\mathcal{E}}}{j} \cdot q_{max}^j \cdot (1 - q_{max})^{N_{\mathcal{E}} - j}. \tag{E2}$$

And as $\binom{N_{\mathcal{E}}}{j} \leq N_{\mathcal{E}}^j$, and $1 - q_{max} \leq 1$, we have that

$$p_{canc} \leq \sum_{j=2}^{N_{\mathcal{E}}} (N_{\mathcal{E}} q_{max})^j. \tag{E3}$$

As $N_{\mathcal{E}} \in \mathbb{Z}^+$, the condition $N_{\mathcal{E}} q_{max} \ll 1$ implies $q_{max} \ll 1$. And so the leading order contribution to this sum is from the cancellation of two errors, i.e. the $j = 2$ term, and the probability that two errors cancel is bounded by $(N_{\mathcal{E}} q_{max})^2$. So that

$$p_{canc} \leq O((N_{\mathcal{E}} q_{max})^2), \tag{E4}$$

This bound will generally be loose, as it is physically reasonable to expect that $p_{canc} \ll O((N_{\mathcal{E}} q_{max})^2)$. And in the regime where $N_{\mathcal{E}} q_{max} \ll 1$, the value of $(N_{\mathcal{E}} q_{max})^2$ will be very small. So that the probability of cancellation is sufficiently small that it can reasonably be neglected, i.e. $O(p_{canc}) \approx 0$. This substitution incurs a small approximation error in the TVD bound. Including the approximation error, $\epsilon_{canc}$, the value used in the TVD computation is then $\tilde{p}_{canc} = 0 + \epsilon_{canc}$. If this expression is used to compute the TVD bound, we have that

$$\frac{2p_{inc}(1 + \epsilon_{canc})}{1 - \epsilon_{canc}^2} \approx 2p_{inc}(1 + \epsilon_{canc}). \tag{E5}$$

Since $p_{inc} \leq 1$, the magnitude of the error in the $\gamma$ bound is similarly bounded $|\epsilon_\gamma| \leq O((N_{\mathcal{E}} q_{max})^2)$. In the regime where $N_{\mathcal{E}} q_{max} \ll 1$, the value of $(N_{\mathcal{E}} q_{max})^2$ will be very small. □

### 2. Bounding the probability of error cancellation for errors affecting solely $\mathcal{J}$-type gate layers in a trap circuit

In the target and trap circuits, the $\mathcal{J}$-type gate layers contain any gates performed by the consumption of magic states. In early fault-tolerant computation most likely those gates performed using imperfectly purified magic states will have logical error rates considerably higher than those of the rest of the circuit operations. In this scenario, the cancellation of errors from the $\mathcal{J}$-type gate layers will be the dominant effect. We now show that the probability of error cancellation due to errors affecting $\mathcal{J}$-type gate layers in the trap circuits can be upper-bounded by a constant. This is formalised in the following lemma.

**Lemma 7.** *The value of the probability of cancellation for errors affecting $\mathcal{J}$-type gate layers in the trap circuits is upper bounded $p_{canc} \leq 1/2$.*

*Proof.* A trap circuit in which two nearest-neighbour $\mathcal{J}$-type gate layers are affected by errors may be expressed in the form

$$\begin{aligned}\mathcal{H}^t \circ (\circ_{j=l+2}^{m} \mathcal{W}_{3j}\mathcal{J}_{3j-1}\mathcal{W}_{3j-2}) \circ (\mathcal{W}_{3(l+1)}P_{3(l+1)-1}\mathcal{J}_{3(l+1)-1}\mathcal{W}_{3(l+1)-2}) \circ (\mathcal{W}_{3l}P_{3l-1}\mathcal{J}_{3l-1}\mathcal{W}_{3l-2})\\ \circ (\circ_{j=1}^{l-1}\mathcal{W}_{3j}\mathcal{J}_{3j-1}\mathcal{W}_{3j-2}) \circ \mathcal{H}^t,\end{aligned} \tag{E6}$$

where $P_{3(l+1)-1}$ is a Pauli error affecting the gate layer $\mathcal{J}_{3(l+1)-1}$, and $P_{3l-1}$ is a Pauli error affecting the gate layer $\mathcal{J}_{3l-1}$. As $\mathcal{W}_{3j}\mathcal{J}_{3j-1}\mathcal{W}_{3j-2} = \mathcal{J}_{3j-1}{}'$ where $\mathcal{J}_{3j-1}{}'$ is a gate layer of randomly oriented CNOT gates and identity operations, and $\mathcal{W}_{3j} = \mathcal{W}_{3j-2}^{\dagger}$ in the trap circuits, we can rewrite the previous expression as

$$\mathcal{H}^t\circ(\circ_{j=l+2}^{m}\mathcal{J}_{3j-1}{}') \circ (\mathcal{W}_{3(l+1)}P_{3(l+1)-1}\mathcal{W}_{3(l+1)-2}\mathcal{J}_{3(l+1)-1}{}') \circ (\mathcal{W}_{3l}P_{3l-1}\mathcal{W}_{3l-2}\mathcal{J}_{3l-1}{}') \circ (\circ_{j=1}^{l-1}\mathcal{J}_{3j-1}{}') \circ \mathcal{H}^t. \tag{E7}$$

Now as conjugation of each of the Pauli errors by $\mathcal{W}$-type gate layers just maps them to other Pauli operators, the previous expression becomes

$$\mathcal{H}^t\circ(\circ_{j=l+2}^{m}\mathcal{J}_{3j-1}{}') \circ P'_{3(l+1)-1} \circ \mathcal{J}_{3(l+1)-1}{}' \circ P'_{3l-1} \circ \mathcal{J}_{3l-1}{}' \circ (\circ_{j=1}^{l-1}\mathcal{J}_{3j-1}{}') \circ \mathcal{H}^t. \tag{E8}$$

where $P'_{3(l+1)-1} = \mathcal{W}_{3(l+1)}P_{3(l+1)-1}\mathcal{W}_{3(l+1)-2}$ and $P'_{3l-1} = \mathcal{W}_{3l}P_{3l-1}\mathcal{W}_{3l-2}$. As the conjugating gates in the $\mathcal{W}$-type gate layers are chosen randomly to be either $S$ or $H$ gates for each logical qubit, this randomises the mapping to the new Pauli operator. Now, only considering the Pauli operators and the dividing gate later, we have expression

$$P'_{3(l+1)-1} \circ \mathcal{J}_{3(l+1)-1}{}' \circ P'_{3l-1}. \tag{E9}$$

The Pauli error $P_{3(j+1)-1}{}'$ can be propagated through the Clifford gate layer $\mathcal{J}_{3(j+1)-1}{}'$, leading to the expression

$$\mathcal{J}_{3(l+1)-1}{}' \circ P''_{3(l+1)-1}P'_{3l-1}. \tag{E10}$$

where $P_{3(j+1)-1}{}'\mathcal{J}_{3(l+1)-1}{}' = \mathcal{J}_{3(l+1)-1}{}'P_{3(j+1)-1}{}''$, and where $P_{3(j+1)-1}{}''$ is another Pauli operator. The Pauli error operators cancel if

$$P''_{3(j+1)-1}P'_{3j-1} = I^{\otimes n}. \tag{E11}$$

In the single-qubit case, the Pauli errors $P_{3(j+1)-1}, P_{3j-1} \in \{X, Y, Z\}$ are mapped to new Pauli operators by the random conjugation by $\mathcal{W}$-type gate layers with probability at least 1/2. The $\mathcal{J}$-type layer is an identity operation and thus does not affect error propagation. In the single-qubit case, the randomisation means that each error can be one of two single-qubit Pauli operators. If the first error is fixed, the randomisation of the second error means that cancellation can happen with, at most, probability 1/2. In the worst case, cancellation occurs for two out of the four possible error combinations, and so

$$\Pr\Big(P''_{3(j+1)-1}P'_{3j-1} = I\Big) \le 1/2. \tag{E12}$$

Similar analysis can be performed for the two-qubit case. Now $P_{3(j+1)-1}, P_{3j-1} \in \{I, X, Y, Z\}^{\otimes 2}\backslash I^{\otimes 2}$, and the $\mathcal{J}$-type layer can be either an $I^{\otimes 2}$ operation or a randomly oriented CNOT gate. The sandwiching gate layers randomly map the Pauli errors to new Pauli errors with probability at least 1/2. And, when moving the Pauli errors together to combine them, the randomly-oriented CNOT gate randomises their propagation, mapping the Pauli errors to new Pauli errors when they are moved through the gate. The random sandwiching layers and the CNOT gate mean that after the two errors have been propagated together to $P''_{3(j+1)-1}P'_{3j-1}$, both $P''_{3(j+1)-1}$ and $P'_{3j-1}$ can, with uniform probability, each be at least two different Pauli errors. In the worst case, both errors are effectively chosen uniformly at random from a set of two Pauli operators, and error cancellation occurs with probability 1/2. It follows that

$$\Pr\Big(P''_{3(j+1)-1}P'_{3j-1} = I^{\otimes 2}\Big) \le 1/2. \tag{E13}$$

For Pauli errors of weight greater than two, the probability of error cancellation may be bounded by bounding the probability that all one or two-qubit components of the errors cancel. For each logical qubit where the $\mathcal{J}$-type gate layer acts with a local identity operation, eqn. E12 may be applied to bound the probability that the components of the errors acting on these qubits cancel. For each pair of logical qubits where the $\mathcal{J}$-type gate layer acts with a randomly oriented CNOT operation, eqn. E13 may be applied to bound the probability that the components of the errors acting on each such pair of qubits cancel. As all such components of the errors must cancel for the errors to cancel overall, the probability of error cancellation is upper-bounded by $(1/2)^{a+b}$, where $a$ is the number qubits for which the component of the $\mathcal{J}$-type layer is identity and $b$ is the number of CNOT operations in the layer. It follows that

$$\Pr\Big(P''_{3(j+1)-1}P'_{3j-1} = I^{\otimes n}\Big) \leq 1/2, \tag{E14}$$

where $n$ is an arbitrary number of logical qubits. If two errors affect $\mathcal{J}$-type gate layers that are not nearest-neighbour, the same bound can be derived using a similar approach, the only difference being that the Pauli errors must be propagated through multiple $\mathcal{J}$-type layers instead of just one. So, any two Pauli errors that affect the $\mathcal{J}$-type gate layers of a randomly generated trap circuit cancel with a probability upper-bounded by 1/2.

We now consider the case where three errors affect the $\mathcal{J}$-type gate layers of a randomly generated trap circuit. We have just shown that the probability of cancellation of any two errors affecting $\mathcal{J}$-type gate layers within a trap circuit is upper-bounded by 1/2, now let us assume the worst-case cancellation probability of $c = 1/2$. If two of the three errors are brought together, they cancel with probability $c$ and lead to a different Pauli error with probability $1 - c$. The probability they combine into a different Pauli error which then cancels with the remaining error is $(1 - c)c$.Therefore, the probability that the three errors cancel is $(1 - c)c = (1/2) \cdot (1/2) \leq c$. Let the notation: $\Pr(j = 2|\text{ cancel})$ indicate the probability that Pauli errors affecting any two $\mathcal{J}$-type gate layers in a trap circuit cancel. Now since $\Pr(j = 2|\text{ cancel}) \leq 1/2$ and $\Pr(j = 3|\text{ cancel}) \leq 1/2$, we have that

$$\begin{aligned}\Pr(j = 4|\text{ cancel}) = (1 - c) \cdot \Pr(j = 3|\text{ cancel})& \\ + c \cdot \Pr(j = 2|\text{ cancel})& \\ \leq (1 - 1/2) \cdot 1/2 + 1/2 \cdot 1/2& \\ \leq 1/2.&\end{aligned} \tag{E15}$$

This argument can be iteratively applied to any number of errors that affect a trap circuit. So that the probability of cancellation for any collection of errors affecting the sequence is upper-bounded by 1/2. Therefore, the probability that any collection of errors affecting the gates performed using magic states in trap circuits cancel is $p_{canc} \leq 1/2$. □

We now prove Lemma 5, that states:

*In the regime where the dominant source of noise is due to the $\mathcal{J}$-type gate layers, the probability of error cancellation is bounded*

$$p_{canc} \leq 1/2 + O(N_{\mathcal{J}} \cdot q_{\mathcal{J},max} \cdot N_{\mathcal{C}} \cdot q_{\mathcal{C},max}), \tag{E16}$$

*where $N_{\mathcal{J}}$ is the number of logical noise channels affecting $\mathcal{J}$-type gate layers in the circuit and $q_{\mathcal{J},max}$ is the highest total channel error rate of these, and $N_{\mathcal{C}}$ is the number of logical noise channels affecting logical single-qubit Clifford gate layers, and state preparation and measurement, and $q_{\mathcal{C},max}$ is the highest of these. In this regime, it may be assumed that $p_{canc} \leq \beta$, where $\beta = 1/2$. If this $\beta$ value is used to compute the TVD upper-bound, $\gamma$, the magnitude of the approximation error from this assumption is bounded*

$$|\epsilon_\gamma| \leq O\Big(N_{\mathcal{J}} \cdot q_{\mathcal{J},max} \cdot N_{\mathcal{C}} \cdot q_{\mathcal{C},max}\Big). \tag{E17}$$

*Proof.* If there are $N_{\mathcal{J}}$ logical noise channels that affect the logical $\mathcal{J}$-type gate layers of each trap circuit, the set of total error rates for these channels is denoted by $\{q_{\mathcal{J},tot}(j)\}_{j=1}^{N_{\mathcal{J}}}$, and the highest error rate from these is $q_{\mathcal{J},max} = \max(\{q_{\mathcal{J},tot}(j)\}_{j=1}^{N_{\mathcal{J}}})$. Likewise, if there are $N_{\mathcal{C}}$ logical noise channels that affect the logical single-qubit Clifford gate layers, and the state preparation and measurement operations, the set of total error rates for these channels is denoted by $\{q_{\mathcal{C},tot}(i)\}_{i=1}^{N_{\mathcal{C}}}$, and the highest error rate from these is $q_{\mathcal{C},max} = \max(\{q_{\mathcal{C},tot}(i)\}_{i=1}^{N_{\mathcal{C}}})$.

Let $\epsilon_{canc}$ denote the approximation error that arises from only considering the cancellation of errors that affect $\mathcal{J}$-type gate layers in the derivation of an upper bound for $p_{canc}$. There are two factors contributing to this approximation error. These are: (1) the cancellation of errors affecting $\mathcal{J}$-type gate layers with errors affecting the rest of the circuit, and (2) the cancellation of any errors excluding those affecting $\mathcal{J}$-type gate layers. The first of these may be bounded by convolving the binomial distributions constructed by using the worst-case channel error rate values of $q_{\mathcal{J},max}$ and $q_{\mathcal{C},max}$ as the binomial probabilities. The second may be bounded by taking the binomial distribution constructed by using the worst-case channel error value $q_{\mathcal{C},max}$ as the binomial probability. So that

$$\begin{aligned}\epsilon_{canc} \le \sum_{j=2}^{N_\mathcal{J}}\sum_{k=1}^{j}\Bigg(\binom{N_\mathcal{J}}{k}\cdot q_{\mathcal{J},max}^{k}\cdot(1-q_{\mathcal{J},max})^{N_\mathcal{J}-k}\Bigg)\cdot\Bigg(\binom{N_\mathcal{C}}{j-k}\cdot q_{\mathcal{C},max}^{j-k}\cdot(1-q_{\mathcal{C},max})^{N_\mathcal{C}-(j-k)}\Bigg)\\ + \sum_{i=2}^{N_\mathcal{C}}\binom{N_\mathcal{C}}{i}\cdot q^{i}_{\mathcal{C},max}\cdot(1-q_{\mathcal{C},max})^{N_\mathcal{C}-i}.\end{aligned} \tag{E18}$$

If the dominant source of errors affecting the circuit is the $\mathcal{J}$-type gate layers, then $q_{\mathcal{J},max} \gg q_{\mathcal{C},max}$. And so, considering only the leading order terms, the previous bound becomes

$$\begin{aligned}\epsilon_{canc} &\le O\Big((N_\mathcal{J}\cdot q_{\mathcal{J},max}\cdot(1-q_{\mathcal{J},max})^{N_\mathcal{J}-1})\\ &\qquad\cdot(N_\mathcal{C}\cdot q_{\mathcal{C},max}\cdot(1-q_{\mathcal{C},max})^{N_\mathcal{C}-1})\Big)\\ &\le O\Big(N_\mathcal{J}\cdot q_{\mathcal{J},max}\cdot N_\mathcal{C}\cdot q_{\mathcal{C},max}\Big).\end{aligned} \tag{E19}$$

Lemma 7 states that errors affecting the $\mathcal{J}$-type gate layers cancel with probability upper bounded by 1/2. Now, including the bound for the approximation error, the probability of error cancellation is bounded by

$$p_{canc} \le 1/2 + O(N_\mathcal{J}\cdot q_{\mathcal{J},max}\cdot N_\mathcal{C}\cdot q_{\mathcal{C},max}). \tag{E20}$$

If we set $p_{canc} = 1/2$, then, including the approximation error from only considering the probability of cancellation of errors from $\mathcal{J}$-type gate layers, if the value $\tilde{p}_{canc} = 1/2 + O(N_\mathcal{J}\cdot q_{\mathcal{J},max}\cdot N_\mathcal{C}\cdot q_{\mathcal{C},max})$ is used to compute the TVD bound, the resulting expression is

$$\begin{aligned}p_{err} &\le \frac{2p_{inc}(1/2+\epsilon_{canc})}{(1/4-\epsilon^2_{canc})}\\ &\approx 8p_{inc}(1/2+\epsilon_{canc}).\end{aligned} \tag{E21}$$

Since $p_{inc} \le 1$, the magnitude of the error in the $\gamma$ bound is similarly bounded $|\epsilon_\gamma| \le O\Big(N_\mathcal{J}\cdot q_{\mathcal{J},max}\cdot N_\mathcal{C}\cdot q_{\mathcal{C},max}\Big)$. In the regime where both $N_\mathcal{J}\cdot q_{\mathcal{J},max} \ll 1$ and $N_\mathcal{C}\cdot q_{\mathcal{C},max} \ll 1$, the value of this upper-bound will be very small. □

### 3. Bounding the probability of error cancellation for any errors affecting a trap circuit

We now describe a *modified trap circuit* construction, and show that this allows the error cancellation probability for errors occurring anywhere in a randomly generated circuit to be upper-bounded by 7/8. Without loss of generality, we will assume magic states are only used to perform $Z$-type rotations in the target and trap circuits, i.e. these logical gate operations are all of the form $Z^{\otimes a}(\theta)$ for $a \in \{1,\ldots,n\}$. This does not restrict computation, since straightforward conjugation using Clifford gates can be used to arbitrarily change the basis of the Pauli rotation.

While otherwise the same as the previous construction, the modified trap circuit construction has three differences; these are:

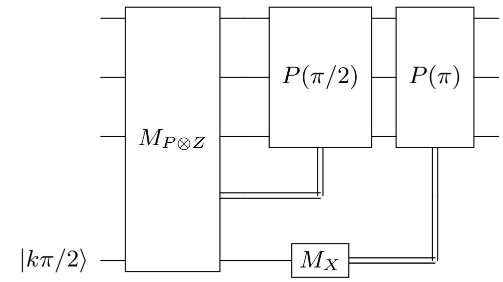


FIG. 13: In the trap construction described in Section E 3, the value of $k$ is randomly chosen where $k \in \{1,2\}$ with probability 1/2. This means that either a $P(\pi/2)$ or $P(\pi)$ Pauli rotation gate is applied, depending on the magic state that is prepared. If a $P(\pi/2)$ rotation is applied, this stabilises the quantum state with probability 1/2, otherwise a correction operation must be included at the end of the circuit to ensure that the trap computation is deterministic. The two adaptive correction operations after the adaptive measurements are applied in the same way as in the target circuits. Namely if the rotation is in the wrong direction the $P(\pi/2)$ operation is applied. So that if a $|\pi/2\rangle$ magic state is used, the overall operation the gadget performs is an identity operation or a $P(\pi/2)$ rotation operation. While if a $|\pi\rangle$ magic state is used, the overall operation the gadget performs is an $P(-\pi/2)$ rotation operation or a $P(\pi)$ rotation operation.

1. The ordering of neighbouring single-qubit Clifford $\mathcal{W}$-type gate layers is randomised.

2. Magic state preparation for the trap circuits is randomised. Either a $|\pi\rangle$ or $|\pi/2\rangle$ magic state is prepared, each with probability 1/2. Consequently, the gates performed using these magic states are rotations of $\pi$ or $\pi/2$, each with probability 1/2. The same correction gates are applied in the same way as if a $|\pi/4\rangle$ magic state was being used to apply a $\pi/4$ rotation by projective measurement in the target circuit; this is shown in Fig. 13. Including the correction gates, the overall gate operations performed are then of the form $Z^{\otimes a}(k\cdot\pi/2)$ for $a \in \{1,\ldots,n\}$ and $k \in \{0,1\}$. The value of $k$ is then 0 or 1 depending on whether the magic state used is $|\pi/2\rangle$ or $|\pi\rangle$, and on the random direction of the initial projective measurement operation.

3. To ensure that the outputs of the trap circuits are deterministic in the absence of error, an adjoint operation is included for each $\pi/2$ rotation gate that does not stabilise the quantum state. These adjoint operations are propagated to the end of each trap circuit, and the order in which they are applied at

the end of the circuit is randomised. Of these operations, those that commute may be parallelised, while those that combine into an identity operation can be removed entirely.

The probability of error cancellation resulting from the additional adjoint operations is later bounded. We expect this result to only be useful in the case of large fault-tolerant circuits, where the number of adjoint operations for every trap would be very close to the mean, and so the error rates of the traps is very similar. We can then assume that the trap error rates are the same, while being larger than the target circuit error rate, and invoke the robustness result (see 2) to argue that any small changes in the error rate will lead to small variations in the computed TVD bound. And so any bound derived using the outputs of these modified trap circuits is a valid upper-bound for the TVD of the target circuit.

Using this construction, each trap circuit is of the form

$$\mathcal{H}^t \circ (\circ_{j=1}^{m} \mathcal{W}_{3j} \mathcal{J}_{3j-1} \mathcal{W}_{3j-2}) \circ (\circ_{k=1}^{p} \mathcal{C}_k) \circ \mathcal{H}^t. \tag{E22}$$

Note that neighbouring $\mathcal{W}$-type gate layers are randomly ordered. And if the ordering of an $S$ and an $H$ gate is reversed by this randomisation, the $S$ gate changes to an $X(\pi/2)$ gate. Also the $(\circ_{k=1}^{p} \mathcal{C}_k)$ gate layers are randomly ordered adjoint operations of the $\pi/2$ $Z$-basis rotations that do not stabilise the state $\mathcal{H}^t \left|0\right\rangle^{\otimes n}$. Where, the same as for the previous construction, the value of $t$ is chosen uniformly at random from the set $\{0, 1\}$.

We now show that any two Pauli errors that occur while performing such a trap circuit cancel with bounded probability.

We now prove Lemma 6, that states:

*For the modified trap circuit construction, the probability of error cancellation in the trap circuits is upper bounded by $p_{canc} \leq 7/8$.*

*Proof.* This proof is organised as follows. We first show that errors that occur at any two positions in a randomly generated trap circuit cancel with bounded probability. We then extend this to bound the error cancellation probability for any number of errors affecting a randomly generated trap circuit.

By Lemma 7, the probability of cancellation for any two errors affecting $J$-type layers is upper-bounded by 1/2. Likewise, using arguments similar to those used in the proof of Lemma 7, any two errors separated by randomly chosen $\mathcal{W}$-type gate layers, where these intervening gate layers do not compile together into an identity operation, are randomly mapped to different Pauli errors when the errors are propagated together and so cancel with probability $\leq 1/2$. This includes the case of a gate preparation or measurement error with an error affecting a gate layer, and the case of errors affecting two $\mathcal{W}$-type gate layers.

For readability, the rest of the proof is structured in sections. In Sections $(a)$-$(e)$, upper-bounds are derived for the probability of cancellation for all other cases where two Pauli errors affect the trap circuit. While in Section $(f)$, the analysis is extended to bound the probability of cancellation for any number of errors.

### *a. Cancellation of state preparation and measurement errors*

A trap circuit affected by state preparation and measurement Pauli errors may be expressed in the form

$$P_M \mathcal{H}^t \circ (\circ_{j=1}^{m} \mathcal{W}_{3j} \mathcal{J}_{3j-1} \mathcal{W}_{3j-2}) \circ (\circ_{k=1}^{p} \mathcal{C}_k) \circ \mathcal{H}^t P_P, \tag{E23}$$

where $P_P$ and $P_M$ denote state preparation and measurement Pauli errors. Firstly, without changing the logic of the trap circuit, the ordering of neighbouring $\mathcal{W}$-type layers is un-randomised to

$$P_M \mathcal{H}^t \circ (\circ_{j=1}^{m} \mathcal{W}'_{3j} \mathcal{J}_{3j-1} \mathcal{W}'_{3j-2}) \circ (\circ_{k=1}^{p} \mathcal{C}_k) \circ \mathcal{H}^t P_P, \tag{E24}$$

where the prime notation indicates the un-randomisation or the gate layer ordering. The gate layers separating the two Pauli errors compile to a mixture of randomly oriented $CNOT$ gates, $X$- or $Z$-type $\pi/2$ rotation gates, and identity operations. The $\left|0\right\rangle^{\otimes n}$ state is always initially prepared, and measurement is always in the $Z$ basis. So the only state preparation and measurement Pauli errors that do not stabilise the initial state or measurement, and non-trivially affect the target and trap computations, are $X$-type errors.

The randomised construction of the trap circuit means that the propagation of state preparation and measurement errors through the trap circuit is randomised so that such errors cancel with bounded probability. This may be seen by propagating the state preparation and measurement Pauli errors through the trap circuit until they are either side of the $j$-th gate layer sequence from $(\circ_{j=1}^{m} \mathcal{W}'_{3j} \mathcal{J}_{3j-1} \mathcal{W}'_{3j-2})$, i.e. $\mathcal{W}'_{3j} \mathcal{J}_{3j-1} \mathcal{W}'_{3j-2}$, where the $\mathcal{J}$-type layer has non-trivial support on at least one of the same qubits as one of the propagated Pauli errors. Such a $J$-type gate layer must exist for the qubits that the errors are affecting to be non-trivially included in the target computation. The errors and the gate layers may then be considered independently of the rest of the circuits as

$$P'_M (\mathcal{W}'_{3j} \mathcal{J}_{3j-1} \mathcal{W}'_{3j-2}) P'_P, \tag{E25}$$

and because the gate layers compile together into randomly oriented CNOT gates, and either $X$-or $Z$-type $\pi/2$ rotations or identity operations, the propagation of the Pauli errors through these intervening gate layers is randomised. Any randomly oriented CNOT in $\mathcal{J}_{3j-1}$ that has support on at least one qubit with at least of the Pauli errors randomises the propagation of the Pauli errors through it, resulting in cancellation of the elements of the Pauli errors with support on the same qubits as the CNOT with probability $\leq 1/2$. While any Pauli rotation operation in $\mathcal{J}_{3j-1}$, and which has support on at least one of the same qubits as at least one of the Pauli errors, is a rotation by $\pi/2$ with probability $1/2$. With probability $1/2$ any Pauli rotation operation stabilises the quantum state, and so does not require an adjoint operation to keep the circuit deterministic in the absence of error, meaning the random propagation of the errors though the rotation operation is not undone by the adjoint operation. Furthermore, the sandwiching $\mathcal{W}$-type gate layers randomise the rotation basis to be either $Z$-type or $X$-type with probability $1/2$. Therefore, a Pauli operation in $\mathcal{J}_{3j-1}$ is: (1) a $\pi/2$ rotation, (2) stabilises the quantum state, and (3) does not commute with the Pauli errors, with probability at least $1/2^3$. This means that the propagation through the circuit of every possible state preparation error and measurement error is randomised with a probability of at least $1/8$. Consequently, the probability of state preparation errors occurring and cancelling in a randomly generated trap circuit is $\leq 7/8$.

### b. *Cancellation of gate errors and state preparation or measurement errors*

Only Pauli errors with a $X$-type component non-trivially affect state preparation and measurement. The same is true of errors affecting the final gate layer of the circuit, and these errors may be combined with measurement errors. Errors affecting the final gate layer of the circuit may be combined with measurement errors, and similarly are non-trivial only when they have an $X$ component. The modified trap construction ensures that any Pauli error affecting a gate layer is separated from errors arising during state preparation or measurement by at least one randomly chosen single-qubit Clifford gate layer. Hence, the propagation required to combine either a state preparation or measurement Pauli error with a gate Pauli error is randomised. And, by similar arguments as were used to prove Lemma 7, any state preparation or measurement Pauli error cancels with a gate Pauli error with probability $\leq 1/2$.

### c. *Cancellation of gate errors occurring before and after J-type layers*

A trap circuit affected by errors occurring before and after a $J$-type layer may be expressed in the form

$$\mathcal{H}^t \circ (\circ_{j=l+1}^{m} \mathcal{W}_{3j}\mathcal{J}_{3j-1}\mathcal{W}_{3j-2}) \circ (\mathcal{W}_{3l} P_{3l-1}\mathcal{J}_{3l-1}P_{3l-1-2}\mathcal{W}_{3l-2}) \circ (\circ_{j=1}^{l-1} \mathcal{W}_{3j}\mathcal{J}_{3j-1}\mathcal{W}_{3j-2}) \circ \mathcal{H}^t. \tag{E26}$$

Isolating the errors and the intervening gate layers from the rest of the circuit, this becomes

$$P_{3l-1}\mathcal{J}_{3l-1}P_{3l-1-2}. \tag{E27}$$

As the $\mathcal{J}$-type layer is a mixture of $CZ$, identity and $Z$-type rotations in the target, and the gate positioning and type is unchanged in the trap circuits, any Pauli errors that cancel in the target circuit also cancel in the trap circuits. Cancellation of these errors in the trap circuits is permitted since these errors do not contribute to the total circuit error rate of the target circuit. Since the rotation angle of the $Z$-type rotations in the traps is chosen to be $\pi/2$ or 0 each with probability $1/2$. This randomises the propagation of errors in the trap circuits that do not cancel in the target circuit with probability $1/2$, and so the probability for errors of this type cancelling in the trap circuits is $\leq 1/2$.

### d. *Cancellation of gate errors occurring between sequences of sandwiched gate layers*

A trap circuit instance affected by gate errors occurring between sequences of sandwiching and $\mathcal{J}$-type errors may be written

$$\mathcal{H}^t \circ (\circ_{j=l+1}^{m} \mathcal{W}_{3j}\mathcal{J}_{3j-1}\mathcal{W}_{3j-2}) \circ P_{3l} \circ (\mathcal{W}_{3l}\mathcal{J}_{3l-1}\mathcal{W}_{3l-2}) \circ P_{3(l-1)} \circ (\circ_{j=1}^{l-1} \mathcal{W}_{3j}\mathcal{J}_{3j-1}\mathcal{W}_{3j-2}) \circ \mathcal{H}^t. \tag{E28}$$

Again isolating the errors and gate layers dividing them from the rest of the circuit, we have

$$P_{3l}(\mathcal{W}_{3l}\mathcal{J}_{3l-1}\mathcal{W}_{3l-2})P_{3(l-1)}. \tag{E29}$$

Now since the ordering of neighbouring Clifford sandwiching layers is randomised, the Pauli errors are randomly mapped to new Pauli operators when propagated together. So that for

$$(\mathcal{W}_{3l}\mathcal{J}_{3l-1}\mathcal{W}_{3l-2})P'_{3l}P_{3(l-1)}, \tag{E30}$$

where $(\mathcal{W}_{3l}\mathcal{J}_{3l-1}\mathcal{W}_{3l-2})P'_{3l} = P_{3l}(\mathcal{W}_{3l}\mathcal{J}_{3l-1}\mathcal{W}_{3l-2})$, the expression $P'_{3l}P_{3(l-1)} = I$ holds with probability $\leq 1/2$. This follows from the randomised ordering of $\mathcal{W}$-type gate layers in the modified trap circuits.

#### *e. Cancellation of any errors with errors from adjoint Clifford gate layers*

In this trap circuit construction, Pauli rotation gates applied by the consumption of magic states randomly perform rotations by either $\pi/2$ or 0. If a rotation by $\pi/2$ is performed, this operation stabilises the quantum state with probability 1/2. This depends on whether the layer of Hadamard gates is applied at the beginning and end of the circuit, making the quantum state either $|0\rangle^{\otimes n}$ or $|+\rangle^{\otimes n}$, and on the choice of random sandwiching gates, rendering the rotation gate basis either $Z$-type or $X$-type. If the gate does not stabilise the state, then a Clifford correction must be applied that performs the adjoint operation of the Pauli rotation gate. These operations are propagated to the end of the circuit and applied in a random order. All of the trap circuit gate operations are Clifford, and so this propagation may be performed efficiently. As stated, the additional Clifford operation is only necessary when both the Pauli operation is a $\pi/2$ rotation, and when this operation does not stabilise the quantum state. As the probability of each of these is 1/2, the probability of a correction operation being necessary is 1/4. So that the probability that an error affects another part of the circuit and cancels due to an error affecting one of the correction operations is $\leq 1/4$.

#### *f. Cancellation of any number of errors*

We have shown that any two errors affecting the trap circuit in this construction cancel with probability $\leq 7/8$. We now show that any number of errors affect a trap circuit cancel with bounded probability.

Let us assume that the probability of any two errors cancelling is the worst-case cancellation probability of $c = 7/8$. First consider the case where three errors affect a trap circuit. If two of the three errors are brought together, they cancel with probability $c$ and lead to a different Pauli error with probability $1-c$. When they combine to a different error, this error cancels with the remaining error with probability $(1-c)c$. Therefore, the probability that the three errors cancel is $(1-c)c = (1/8)\cdot(7/8) \leq c$. Now let $\mathrm{p}(j = 2|\text{ cancel})$ indicate the probability that any two Pauli errors affecting a trap circuit cancel. Since $\Pr(j = 2|\text{ cancel}) \leq 7/8$ and $\Pr(j = 3|\text{ cancel}) \leq 7/8$, we have that

$$\begin{aligned}\Pr(j = 4|\text{ cancel}) = (1-c)\cdot\Pr(j = 3|\text{ cancel})\\ + c\cdot\Pr(j = 2|\text{ cancel})\\ \leq (1-7/8)\cdot 7/8 + 7/8\cdot 7/8\\ \leq 7/8.\end{aligned} \tag{E31}$$

This argument can be iteratively applied to any number of errors affecting a trap circuit. Therefore, the probability of cancellation for any collection of errors affecting one of the randomly generated modified trap circuits is $p_{canc} \leq 7/8$. □

## Appendix F: Numerical decomposition of false positive rates

We perform numerical analysis on the false positive rates arising in our accreditation protocol applied to IQP circuits under a local logical depolarizing noise model. False positives can be attributed to two main sources. The first is stabilization-induced errors, which correspond to Pauli errors that commute with the trap measurement basis and therefore leave the measurement outcome unchanged, causing false positives. The second source is cancellation-induced errors, which arise from non-commuting errors that, due to circuit symmetries or error propagation, combine to produce a trivial overall effect and thus also result in false positives.

Accordingly, the total false positive probability decomposes as

$$\Pr[\text{false positive}] = p_{\text{stab}} + p_{\text{canc}},$$

where $p_{\text{stab}}$ and $p_{\text{canc}}$ denote the contributions from stabilization and cancellation effects, respectively.

Under the local logical depolarizing noise model (assuming a physical error rate of $p_{phys}$), each error location applies no error with probability $1-p$, or one of the three Pauli errors $\{X, Y, Z\}$ each with probability $p/3$. Given that the trap measurement basis is chosen uniformly at random to be $Z$ or $X$, the per-location probability that an error commutes with the measurement operator is

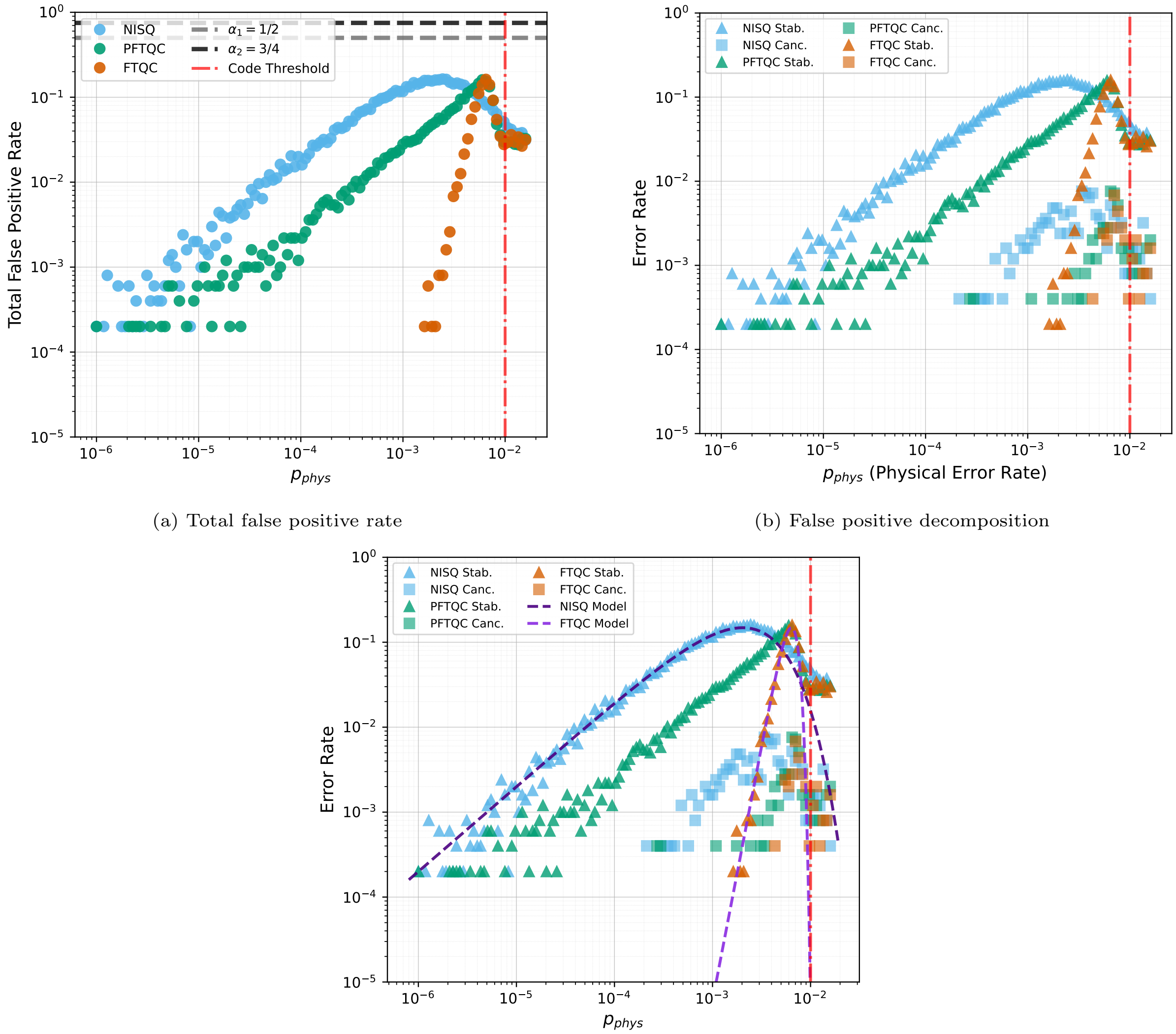
$$p_{\text{stab,loc}} = 1 - \frac{2p}{3}.$$

Assuming $k$ independent error locations, the probability that all errors commute with the trap measurement operator (including the trivial no-error case) is

$$\left(1 - \frac{2p}{3}\right)^k.$$

Subtracting the no-error probability $(1-p)^k$ isolates the false positives arising from nontrivial stabilizing errors:

$$p_{\text{stab}} = \left(1 - \frac{2p}{3}\right)^k - (1-p)^k.$$



(a) Total false positive rate

(b) False positive decomposition

(c) Stabilisation-induced false positives vs. analytical prediction

FIG. 14: *Analysis of false positives in our logical accreditation protocol:* We study the false positive rate in our accreditation protocol for 5-qubit, 40-layer IQP circuits under varying physical error rates. Panel (b) illustrates the two primary sources of false positives: stabilization errors and error cancellation. We find here that the false positives arising from stabilization errors are the dominant contribution to the total false positive rate. In contrast, error cancellation-induced false positives are consistently 1–2 orders of magnitude smaller. In panel (c), we show that for our simplified model assuming local logical depolarizing noise affecting our circuits, the false positive rate from stabilizing errors is well-approximated (for NISQ and FTQC circuits) by the analytical expression $s(k, p_L) = \left(1 - \frac{2p_L}{3}\right)^k - (1 - p_L)^k$, where $k \approx 3 \times \texttt{num\_layers} \times \texttt{num\_qubits}$ denotes the total number of error locations in the circuit.

Our numerical results, shown in Fig. 14(b), confirm that the false positive contribution from cancellation effects is minimal and consistently 1–2 orders of magnitude lower than the contribution from stabilization effects across the full range of physical error rates studied.

These findings support the conclusion that, for stochastic noise models relevant to fault-tolerant quantum computing that are well-approximated by depolarising noise, false positives in our accreditation protocol applied to IQP computations are dominated by stabilization-induced errors. The analytical formula above provides a reliable approximation for this dominant contribution.

**Remark:** *This analysis is primarily valid under the local depolarizing noise model assumed here. However, the implication that stabilization effects dominate the false positive rate likely extends to more general stochastic error models, particularly those where weight-1 errors are the most probable and higher-weight errors occur with decreasing rates. This would generally be true in regimes of fault tolerance, and where the error channels are independent.*

### Appendix G: Robustness bound

Assumptions A1 and A3 can be violated if the logical gate noise channels exhibit logical single-qubit Clifford gate- or physical single-qubit gate-dependence during the protocol. This can result in the logical noise channels affecting the trap circuits differing from those affecting the target circuit. We will now show that the resulting difference in the computed TVD upper-bound when these assumptions are violated depends only linearly on the diamond norm distance of each of the noise channels.

**Definition G.1** (Diamond norm distance)**.** For any two CPTP maps $A$ and $B$ acting on density matrices of a $2^n$-dimensional Hilbert space $\mathcal{H}$, their diamond norm distance is defined as

$$||A - B||_\diamond := \max_\rho(||(A \otimes \mathbb{I})(\rho) - (B \otimes \mathbb{I})(\rho)||_1), \quad \text{(G1)}$$

where $\mathbb{I}$ is identity channel acting on auxilliary system $\mathcal{H}'$ of size $\dim(\mathcal{H}') = \dim(\mathcal{H})$, $||.||_1$ is the $l_1$-norm or trace norm, where $||X||_1 := \mathrm{Tr}(\sqrt{X^\dagger X})$, the density matrix $\rho$ ranges over the Hilbert space $\mathcal{H} \otimes \mathcal{H}'$, and the expression is maximised over all possible density matrices.

In order to derive this result, we will make use of the following lemma:

**Lemma 8.** *For CPTP maps $A$, $B$, $C$ and $D$, the diamond norm, indicated $||\,.\,||_\diamond$, exhibits the property*

$$||AB - CD||_\diamond \leq ||A - C||_\diamond + ||B - D||_\diamond.$$

**Proof**

$$\begin{aligned} ||AB - CD||_\diamond &= ||AB + AD - AD - CD||_\diamond \\ &\leq ||AB - AD||_\diamond + ||AD - CD||_\diamond \\ &\leq ||A|_\diamond \cdot ||B - D||_\diamond + ||A - C||_\diamond \cdot ||D||_\diamond \\ &\leq ||B - D||_\diamond + ||A - C||_\diamond. \end{aligned} \quad \text{(G2)}$$

where the second line follows from a triangle inequality, the third line from the submultiplicity of the diamond norm, and the final line from the property that $||G||_\diamond \leq 1$ for any quantum channel $G$.

#### 1. Proof of Theorem 2

Let the $k$-th logical trap circuit affected by noise, where assumptions A1-A3 hold, be denoted

$$\tilde{\mathcal{C}}^{(k)} = \mathcal{E}_M^{(k)} \cdot \prod_{j=1}^{D} (\mathcal{E}_j^{(k)} \mathcal{L}_j^{(k)}) \cdot \mathcal{E}_P^{(k)}, \quad \text{(G3)}$$

and the $k$-th logical trap circuit affected by noise where assumptions A1-A3 are violated be denoted

$$\tilde{\mathcal{C}}^{(k)\prime} = \mathcal{E}_M^{(k)\prime} \cdot \prod_{j=1}^{D} (\mathcal{E}_j^{(k)\prime} \mathcal{L}_j^{(k)}) \cdot \mathcal{E}_P^{(k)\prime}. \quad \text{(G4)}$$

Applying Lemma 8 to the quantity $||\tilde{\mathcal{C}}^{(k)} - \tilde{\mathcal{C}}^{(k)\prime}||_\diamond$, we have that

$$\begin{aligned} &||\tilde{\mathcal{C}}^{(k)} - \tilde{\mathcal{C}}^{(k)\prime}||_\diamond \\ &\quad = ||\mathcal{E}_M^{(k)} \cdot \prod_{j=1}^{D} (\mathcal{E}_j^{(k)} \mathcal{L}_j^{(k)}) \cdot \mathcal{E}_P^{(k)} - \mathcal{E}_M^{(k)\prime} \cdot \prod_{j=1}^{D} (\mathcal{E}_j^{(k)\prime} \mathcal{L}_j^{(k)}) \cdot \mathcal{E}_P^{(k)\prime}||_\diamond \\ &\quad \leq || \prod_{j=1}^{D} (\mathcal{E}_j^{(k)} \mathcal{L}_j^{(k)}) \cdot \mathcal{E}_P^{(k)} - \prod_{j=1}^{D} (\mathcal{E}_j^{(k)\prime} \mathcal{L}_j^{(k)}) \cdot \mathcal{E}_P^{(k)\prime}||_\diamond + ||\mathcal{E}_M^{(k)} - \mathcal{E}_M^{(k)\prime}||_\diamond. \end{aligned} \quad \text{(G5)}$$

Now this operation can be iteratively applied $2D$ times, resulting in the bound

$$\begin{aligned}
&||\prod_{j=1}^{D}(\mathcal{E}_j^{(k)}\mathcal{L}_j^{(k)})\cdot\mathcal{E}_P^{(k)} - \prod_{j=1}^{D}(\mathcal{E}_j^{(k)\prime}\mathcal{L}_j^{(k)})\cdot\mathcal{E}_P^{(k)\prime}||_\diamond + ||\mathcal{E}_M^{(k)}-\mathcal{E}_M^{(k)\prime}||_\diamond \\
&\quad \leq ||\mathcal{E}_P^{(k)}-\mathcal{E}_P^{(k)\prime}||_\diamond + \sum_{j=1}^{D}||\mathcal{E}_j^{(k)}-\mathcal{E}_j^{(k)\prime}||_\diamond + ||\mathcal{E}_M^{(k)}-\mathcal{E}_M^{(k)\prime}||_\diamond \qquad \text{(G6)} \\
&\quad \leq \sum_{j=0}^{D+1}||\mathcal{E}_j^{(k)}-\mathcal{E}_j^{(k)\prime}||_\diamond,
\end{aligned}$$

where to get the final line, the logical state preparation and measurement noise channels are relabelled using indices '0' and '$D+1$' respectively. And so the diamond norm distance of the $k$-th trap circuit affected by noise where assumptions A1-A3 hold, $\tilde{\mathcal{C}}^{(k)}$, from the $k$-th trap circuit affected by noise where assumptions A1-A3 are violated, $\tilde{\mathcal{C}}^{(k)\prime}$ is bounded

$$||\tilde{\mathcal{C}}^{(k)}-\tilde{\mathcal{C}}^{(k)\prime}||_\diamond \leq \sum_j ||\mathcal{E}_j^{(k)}-\mathcal{E}_j^{(k)\prime}||_\diamond. \qquad \text{(G7)}$$

Now, from the definition of the diamond norm

$$|\rho_{\tilde{\mathcal{C}}^{(k)}}-\rho_{\tilde{\mathcal{C}}^{(k)\prime}}|_1 \leq ||\tilde{\mathcal{C}}^{(k)}-\tilde{\mathcal{C}}^{(k)\prime}||_\diamond,$$

where $\rho_{\tilde{\mathcal{C}}^{(k)}}$ is the output state generated by applying the circuit $\tilde{\mathcal{C}}^{(k)}$ to any input state, and $\rho_{\tilde{\mathcal{C}}^{(k)\prime}}$ is the output state generated by applying the circuit $\tilde{\mathcal{C}}^{(k)\prime}$ to the same input state. Additionally, we have that

$$\delta(\mathcal{D}(\rho_{\tilde{\mathcal{C}}^{(k)}}),\mathcal{D}(\rho_{\tilde{\mathcal{C}}^{(k)\prime}})) \leq |\rho_{\tilde{\mathcal{C}}^{(k)}}-\rho_{\tilde{\mathcal{C}}^{(k)\prime}}|_1,$$

where $\mathcal{D}(.)$ indicates the output probability distribution generated from measuring a quantum state in a given basis. Therefore, it follows that

$$\delta(\mathcal{D}(\rho_{\tilde{\mathcal{C}}^{(k)}}),\mathcal{D}(\rho_{\tilde{\mathcal{C}}^{(k)\prime}})) \leq \sum_j ||\mathcal{E}_j^{(k)}-\mathcal{E}_j^{(k)\prime}||_\diamond.$$

The protocol bound computed if assumptions are violated is of the form $\gamma' = \gamma + \epsilon_d$, where $\gamma$ is the bound where the assumptions hold, $\gamma'$ is the bound computed where the assumptions are violated, and $\epsilon_d$ is the difference in the protocol bound induced by the violation of the assumption. The absolute value of $\epsilon_d$ may be bounded by taking the sum over the eqn. G7 bounds for each of the trap circuits. This provides the robustness bound for the protocol

$$|\gamma-\gamma'| \leq M^{-1}\sum_k\sum_j ||\mathcal{E}_j^{(k)}-\mathcal{E}_j^{(k)\prime}||_\diamond,$$

where the $M^{-1}$ prefactor comes from the fact that the computed upper-bound is proportional to the average of the total circuit error rates of each of the trap circuits, and each of these may have their own noise channel gate- and time-dependence behaviour.

## Appendix H: Upper-bound for the second-order Rényi entropy density using logical accreditation

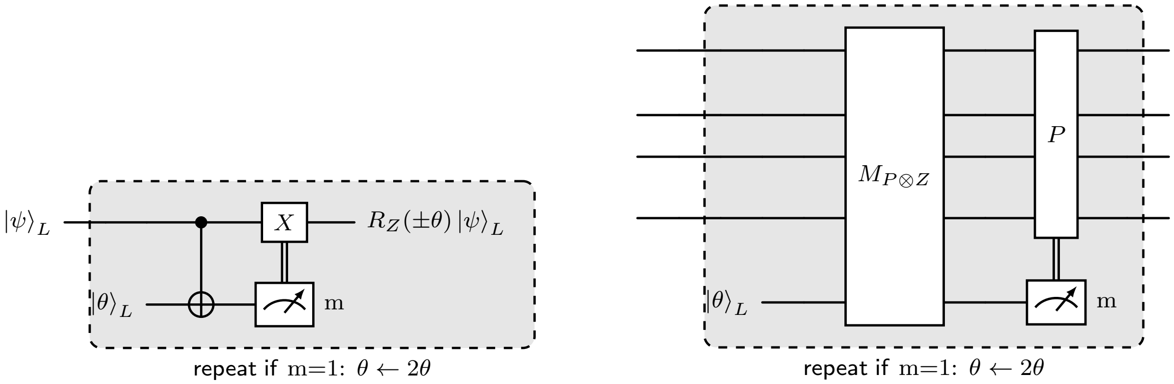
In the logical accreditation protocol, the $n$ computational logical qubit output state of the $i$-th circuit may be expressed in the form

$$\rho_{out}^{(i)} = (1-p_{err})\rho_{out,id}^{(i)} + p_{err}\rho_{noisy}^{(i)}, \qquad \text{(H1)}$$

where $p_{err}$ is the total logical error rate of the logical circuit. The purity of the $i$-th logical output state is then

$$\begin{aligned}
\mathrm{Tr}\big[\rho_{out}^{(i)\,2}\big] &= \mathrm{Tr}\big[\big((1-p_{err})\rho_{out,id}^{(i)} + p_{err}\rho_{noisy}^{(i)}\big)^2\big] && \text{(H2)}\\
&= (1-p_{err})^2 + 2(1-p_{err})p_{err}\mathrm{Tr}(\rho_{out,id}^{(i)}\rho_{noisy}^{(i)}) + p_{err}^2\mathrm{Tr}(\rho_{noisy}^{(i)\,2}) && \text{(H3)}\\
&\geq (1-p_{err})^2 + 2(1-p_{err})p_{err}\mathrm{Tr}(\rho_{out,id}^{(i)}\rho_{noisy}^{(i)}) + p_{err}^2\mathrm{Tr}((\mathbb{1}/2^{2n})^2) && \text{(H4)}\\
&\geq (1-p_{err})^2 + p_{err}^2\mathrm{Tr}(\mathbb{1}/2^{2n}) && \text{(H5)}\\
&\geq (1-p_{err})^2 + p_{err}^2/2^{2n} && \text{(H6)}\\
&\geq 1-2p_{err}+p_{err}^2(1+2^{-n}) && \text{(H7)}\\
&\geq 1-2\gamma+\gamma^2(1+2^{-n}). && \text{(H8)}\\
& && \text{(H9)}
\end{aligned}$$

The distributivity of the trace is used to reach the second line. The inequality of the third line is reached by replacing $\rho_{noisy}^{(i)}$ in the final term with the maximally mixed


$|\psi\rangle_L$
$X$
$R_Z(\pm\theta)\,|\psi\rangle_L$
$|\theta\rangle_L$
m
repeat if m=1: $\theta \leftarrow 2\theta$
$M_{P\otimes Z}$
$P$
$|\theta\rangle_L$
m
repeat if m=1: $\theta \leftarrow 2\theta$

(a)

(b)

FIG. 15: *Repeat-until-success gadgets to implement single- and multi-qubit Pauli rotations using magic states.* (a) A quantum circuit for implementing an analog $Z$ rotation gate $R_Z(\theta)$ where the measurement is a destructive $Z$ basis measurement. (b) A quantum circuit for implementing an analog multi-Pauli rotation gate $R_P(\theta)$, where $P$ is any Pauli string operator. If the measurement outcome of $P \otimes Z$ is 1, i.e. the measured eigenvalue is $-1$, the RUS gate must be repeated. A destructive $X$ basis measurement is used to check whether the output state has a byproduct Pauli operator $P$ that needs to be corrected.

state, using that the maximally mixed state is the maximally entropic state. To obtain the fourth line, we use the observation that

$$\begin{aligned}\mathrm{Tr}(\rho^{(i)}_{out,id}\rho^{(i)}_{noisy}) &= \mathrm{Tr}\big(\,|\psi\rangle\langle\psi|\sum_j \mu_j\,|\phi_j\rangle\langle\phi_j|\,\big)\\ &= \sum_j \mu_j |\,\langle\psi|\phi_j\rangle\,|^2 \qquad \text{(H10)}\\ &\geq 0,\end{aligned}$$

where the density matrices have been initially decomposed in their eigenbases, and the final inequality follows from the positivity of the eigenvalues and of the absolute inner product. The inequality of the final line is reached by replacing $p_{err}$ with the experimentally estimated $2p_{inc}$, this applies with the same confidence as the logical accreditation protocol. The second-order Rényi entropy density of state, $\rho^{(i)}_{out}$, is defined as

$$n^{-1}S^{(2)}(\rho^{(i)}_{out}) = -n^{-1}\log_2(\mathrm{Tr}[\rho^{(i)}_{out}{}^2]). \qquad \text{(H11)}$$

Now, from the final inequality of eqn. H2, and using that for positive reals $x$ and $y$, if $x > y$ then $\log_2(x) > \log_2(y)$, it follows that

$$\log_2\big(\mathrm{Tr}[\rho^{(i)}_{out}{}^2]\big) \geq \log_2(1-2\gamma+\gamma^2(1+2^{-n})). \qquad \text{(H12)}$$

Meaning the second-order Rényi entropy of the logical output state can be upper-bounded using the experimentally estimated output value of $\gamma$, specifically

$$n^{-1}S^{(2)}(\rho^{(i)}_{out}) \leq -n^{-1}\log_2(1-2\gamma+\gamma^2(1+2^{-n})). \qquad \text{(H13)}$$

This bound holds with the same confidence as the logical accreditation bound.

## Appendix I: Frameworks for logical computation

We now describe two frameworks for computation with different logical gate sets, and discuss the application of the certification protocol within each framework. These are: 1) the Clifford and $T$ gate framework, and 2) the Clifford and analog rotation gate framework.

### 1. Clifford and $T$ gate logical circuits

Gate sets consisting of a subgroup of the Clifford group along with the $T$ gate are often considered in logical computation. In the Clifford and $T$ framework, an input computation is compiled using gates from the Clifford group and $T$ gates The inclusion of the $T$ gate is necessary to extend the computational power of the Clifford group for universal computation. Many CSS codes, such as the family of 2D colour codes, can perform Clifford gates transversally, with the non-Clifford $T$ gates needing to be applied through the use of magic state purification and gate teleportation. The Clifford and $T$ framework is a fully fault-tolerant framework, where Clifford gates may be performed using transversal operations or lattice surgery, and $T$ gates are performed using purified magic states. In the Clifford and noisy $T$ gate framework, magic states are not purified in order to avoid the associated space-time overhead [1]. Instead, noisy $|\pi/4\rangle$ magic states are prepared, and are directly used to perform $T$ gates. For computation using either of these frameworks to be compatible with the certification protocol, it is required that any input target computation is compiled using gates from the allowed Clifford gate set $\{CZ, H, S, I, X, Y, Z\}$ and the $T$ gate. Where the Clifford gates may be implemented fault-tolerantly without the use of magic states, and the $T$ gates are implemented by the preparation of $|\pi/4\rangle$ magic states which are then used to perform the $T$ gates by gate teleportation. The logical qubits are required to be initialised in the $|0\rangle^{\otimes n}$ state, and the final logical measurement is in the computational basis. Such a set-up is universal for quantum computation.

### 2. Clifford and analog $Z$-rotation gate logical circuits

Another framework suited to partially fault-tolerant computation uses a gate set of Clifford operations and analog rotation operations. For computation using this framework to be compatible with the certification protocol, it is required that any input target computation is compiled using gates from the allowed Clifford gate set $\{CZ, H, S, I, X, Y, Z\}$ and gates of the form $R_P(\theta)$, where $P \in \{I, X, Y, Z\}^{\otimes n}$ with $n$ the number of logical computation qubits acted on by the rotation. The Clifford gates may be implemented fault-tolerantly without the use of magic states, and the $R_Z(\theta)$ gates by the preparation of noisy $|\theta\rangle$ magic states, which are then used to perform the analog rotations by projective measurement. The logical qubits are required to be initialised in the $|0\rangle^{\otimes n}$ state and the final logical measurement is in the computational basis, this set-up is universal for quantum computation.

One example of a Clifford and analog rotation framework for logical computation is a scheme for partially fault-tolerant logical computation called the space-time efficient analog rotation (STAR) framework [26, 29]. This allows for fault-tolerant Clifford operations and noisy arbitrary Pauli rotations using the surface code. These analog rotations can be local Pauli-$Z$ rotations, shown in Fig. 15 (a), or global logical Pauli rotations as shown in Fig. 15 (b).

## Appendix J: Certifying Trotterised quantum Hamiltonian simulation circuits

Quantum states evolve in time according to the Schrödinger equation: $i\frac{\partial}{\partial t}|\psi(t)\rangle = H(t)|\psi(t)\rangle$, where $\hbar$ is assumed to be 1, $H(t)$ is a possibly time-dependent Hamiltonian, and $|\psi(t)\rangle$ is the quantum state. If the Hamiltonian is time-independent, the exact solution to the Schrodinger equation is: $|\psi(t)\rangle = e^{-iHt}|\psi(0)\rangle$. Trotterization describes the set of methods that apply Trotter-Suzuki decompositions to approximately decompose the exponentially large time evolution operator, $e^{-iHt}$, into a product of smaller matrix exponentials. A Hamiltonian, $H$, that has been mapped to qubits, by, for example, the Jordan-Wigner transformation [58, 59], may be expressed as a linear combination of Pauli operators

$$H = \sum_{i=1}^{L} a_i P_i, \tag{J1}$$

where $P_i \in \{I, X, Y, Z\}^{\otimes n}$ for $n$ computational logical qubits. The Trotter-Suzuki decomposition allows the time evolution operator $e^{-iHt}$ to be approximately decomposed into a sequence of $N$ discrete time-steps. The time evolution operator may be written as a product of $N$ terms

$$e^{-iHt} = \left(e^{-i(\sum_{i=1}^{L} a_i P_i)t/N}\right)^N, \tag{J2}$$

where the evolution operator within the brackets is applied $N$ times and each evolution step is by time $t/N$. The second-order Trotter-Suzuki decomposition approximates these time-step terms as

$$\left(e^{-i(\sum_{i=1}^{L} a_i P_i)t/N}\right)^N \approx \left(\prod_{i=1}^{L} e^{-i(\frac{a_i t}{2N})P_i} \prod_{i=L}^{1} e^{-i(\frac{a_i t}{2N})P_i}\right)^N. \tag{J3}$$

The approximation error of the Trotterised evolution is

$$\left\| e^{-iHt} - \left(\prod_{i=1}^{L} e^{-i(\frac{a_i t}{2N})P_i} \prod_{i=L}^{1} e^{-i(\frac{a_i t}{2N})P_i}\right)^N \right\| \leq \mathcal{O}(t^3), \tag{J4}$$

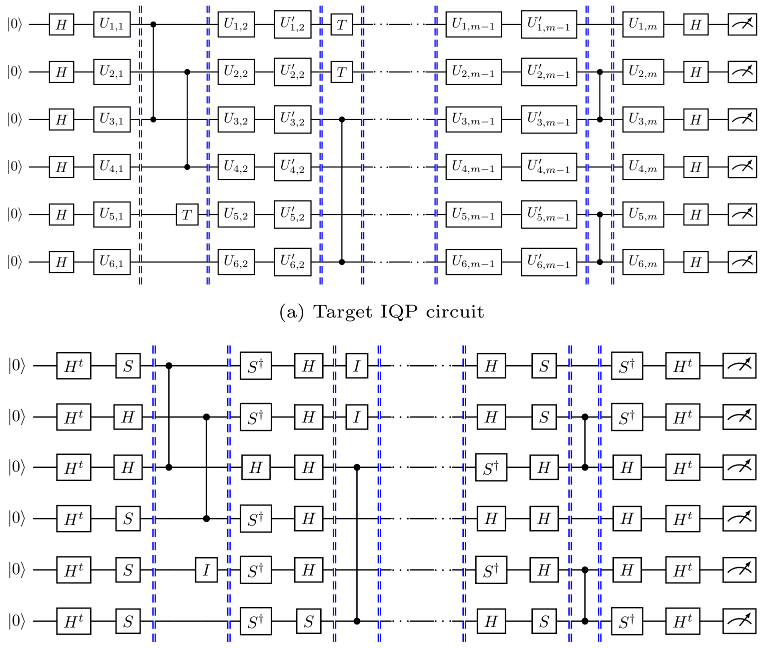
where $\|.\|$ is the operator norm, also known as the spectral norm. Specifically, the upper-bound is $Wt^3$ where $W$ is the Trotter error norm - a constant depending on the Hamiltonian and the type of Trotterization used [45, 47]. As the Hamiltonian is expressed as a linear combination of Pauli operators, the previous decomposition may be written as a sequence of multi-qubit Pauli rotations

$$\left(\prod_{i=1}^{L} e^{-i(\frac{a_i t}{2N})P_i} \prod_{i=L}^{1} e^{-i(\frac{a_i t}{2N})P_i}\right)^N = \left(\prod_{i=1}^{L} R_{P_i}\Big(\frac{a_i t}{2N}\Big) \prod_{i=L}^{1} R_{P_i}\Big(\frac{a_i t}{2N}\Big)\right)^N. \tag{J5}$$

To perform this sequence of operations with a logical circuit, each rotation gate is performed by the consumption of a single magic state. As the logical accreditation protocol permits target circuits including multi-qubit Pauli rotation operations (the construction required is described in Section A) it can be used to provide an upper-bound on the total circuit error of the circuit consisting of a sequence of multi-qubit Pauli operations like eqn. 10, and also an upper-bound on the TVD of the output when measuring any observable of the time-evolved state. The protocol may therefore be used to certify the Trotterised evolution of a Hamiltonian using a logical circuit.

## Appendix K: IQP circuit trap construction

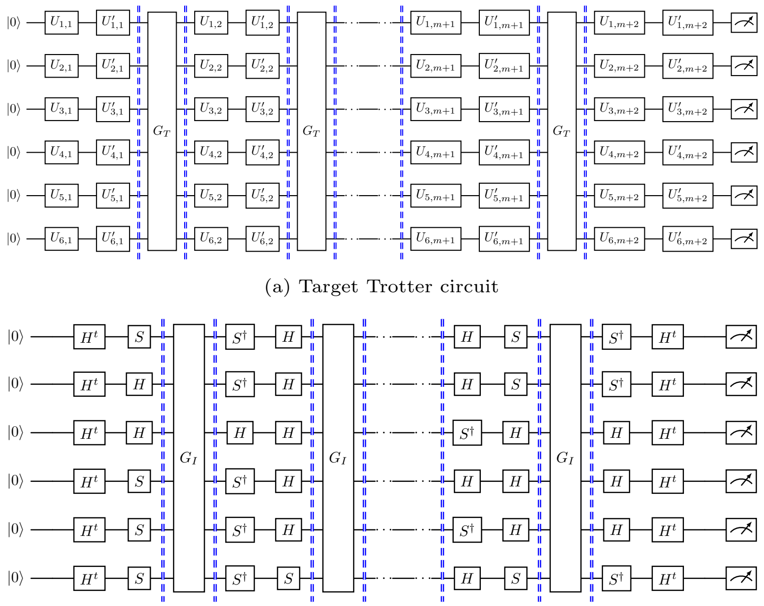
The structure of IQP circuits used for the simulations is shown in Fig. 16, similar to those described in [16]. For the target circuits, layers of single-qubit Clifford phase gates $U_{i,j}$ were alternated with layers of $CZ$ gates and



(a) Target IQP circuit

(b) Trap IQP circuit

FIG. 16: *Structure of target and trap circuits for IQP simulations.* The target IQP circuit (left) and a corresponding trap circuit (right), both using the same layer structure with randomised elements. In the traps, every $CZ$ gate is modified by sandwiching it by $S - S^\dagger$ gates and $H - H$ gates (thus compiling the resulting block of gates into a randomly oriented $CNOT$). For qubits not involved in a $CZ$, we randomly decide between either sandwiching $H - H$ or $S - S^\dagger$ gates.



(a) Target Trotter circuit

(b) Trap Trotter circuit

FIG. 17: *Structure of target and trap circuits for the Trotterised circuit simulation numerical experiments.* In (left) the target Trotterised circuit, and in (right) an example of a trap circuit generated with the same overall structure as the target circuit. The gate layers denoted $G_T$ and $G_I$ are the Trotterised logical circuit operations used for the target and trap circuits, respectively

$T$ gates. In the experiments, the IQP trap circuits were constructed as follows:

- Every $T$ gate was replaced with an identity operation ($I$).
- Each $CZ$ gate was modified by sandwiching it with $S, S^\dagger$, and $H$ gates, as shown in Fig. 11.

- Random $S - S^\dagger$ or $H - H$ gate pairs were applied to qubits not involved in $CZ$ gates.
- For each trap circuit, it was randomly chosen whether to prepend and append Hadamard gate layers (denoted by $H^t$ in Fig. 16) to the circuit.

Finally, in Fig. 18, we compare the total circuit error rate, the TVD upper bound for 5-qubit, 40-layer IQP circuits, using a similar setting as throughout (e.g., distance-11 surface codes for the fault-tolerant configurations). Interestingly, the results suggest that the TVD upper bound not only upperbounds the total circuit error rate (computed here by summing over all depolarising noise channels), but also upperbounds the infidelity between the noisy and ideal output states. We prove this more formally in Appendix L

## Appendix L: Upper-bound for the infidelity between the experimental and ideal output states

The fidelity between the ideal output target circuit state, $\rho^{(0)}_{out,id}$, and the experimental output target state, $\rho^{(0)}_{out}$, is defined by the expression

$$F(\rho^{(0)}_{out,id}, \rho^{(0)}_{out}) = \left(\mathrm{Tr}\sqrt{\sqrt{\rho^{(0)}_{out,id}}\rho^{(0)}_{out}\sqrt{\rho^{(0)}_{out,id}}}\right)^2. \quad \text{(L1)}$$

Since $\rho^{(0)}_{out,id}$ is a pure state, we have that $\rho^{(0)}_{out,id} = |\psi\rangle\langle\psi|$. We know that projection operators are idempotent, so $(|\psi\rangle\langle\psi|)^2 = |\psi\rangle\langle\psi|$. Therefore, $|\psi\rangle\langle\psi| = \sqrt{|\psi\rangle\langle\psi|}$. So the expression for the fidelity then becomes

$$\begin{aligned}\left(\mathrm{Tr}\sqrt{|\psi\rangle\langle\psi|\,\rho^{(0)}_{out}\,|\psi\rangle\langle\psi|}\right)^2 &= \langle\psi|\,\rho^{(0)}_{out}\,|\psi\rangle\left(\mathrm{Tr}\sqrt{|\psi\rangle\langle\psi|}\right)^2\\ &= \langle\psi|\,\rho^{(0)}_{out}\,|\psi\rangle\\ &= \mathrm{Tr}\big(\,|\psi\rangle\langle\psi|\,\rho^{(0)}_{out}\big).\end{aligned} \quad \text{(L2)}$$

The experimental output state $\rho^{(0)}_{out}$ is of the form

$$\rho^{(0)}_{out} = (1 - p_{err})\rho^{(0)}_{out,id} + p_{err}\rho^{(0)}_{noisy}. \quad \text{(L3)}$$

The fidelity of the experimental output state and the ideal state is, therefore,

$$\begin{aligned}&\mathrm{Tr}\Big(\rho^{(0)}_{out,id}\big((1 - p_{err})\rho^{(0)}_{out,id} + p_{err}\rho^{(0)}_{noisy}\big)\Big)\\ &\quad = \mathrm{Tr}\big((1 - p_{err}){\rho^{(0)}_{out,id}}^2 + p_{err}\rho^{(0)}_{noisy}\rho^{(i)}_{out,id}\big)\\ &\quad = (1 - p_{err}) + p_{err}\mathrm{Tr}\big(\rho^{(0)}_{noisy}\rho^{(0)}_{out,id}\big).\end{aligned} \quad \text{(L4)}$$

The infidelity is the complement of the fidelity, so the infidelity between $\rho^{(0)}_{out,id}$ and $\rho^{(0)}_{out}$ is

$$1 - \mathrm{Tr}\Big(\rho^{(0)}_{out,id}\big((1 - p_{err})\rho^{(0)}_{out,id} + p_{err}\rho^{(0)}_{noisy}\big)\Big). \quad \text{(L5)}$$

This may be upper-bounded

$$\begin{aligned}1 - &\mathrm{Tr}\Big(\rho^{(0)}_{out,id}\big((1 - p_{err})\rho^{(0)}_{out,id} + p_{err}\rho^{(0)}_{noisy}\big)\Big)\\ &= 1 - (1 - p_{err}) - p_{err}\mathrm{Tr}\big(\rho^{(0)}_{noisy}\rho^{(0)}_{out,id}\big)\\ &= p_{err}\Big(1 - \mathrm{Tr}\big(\rho^{(0)}_{noisy}\rho^{(0)}_{out,id}\big)\Big)\\ &\leq p_{err}\\ &\leq \gamma.\end{aligned} \quad \text{(L6)}$$

Where we have used the inequality

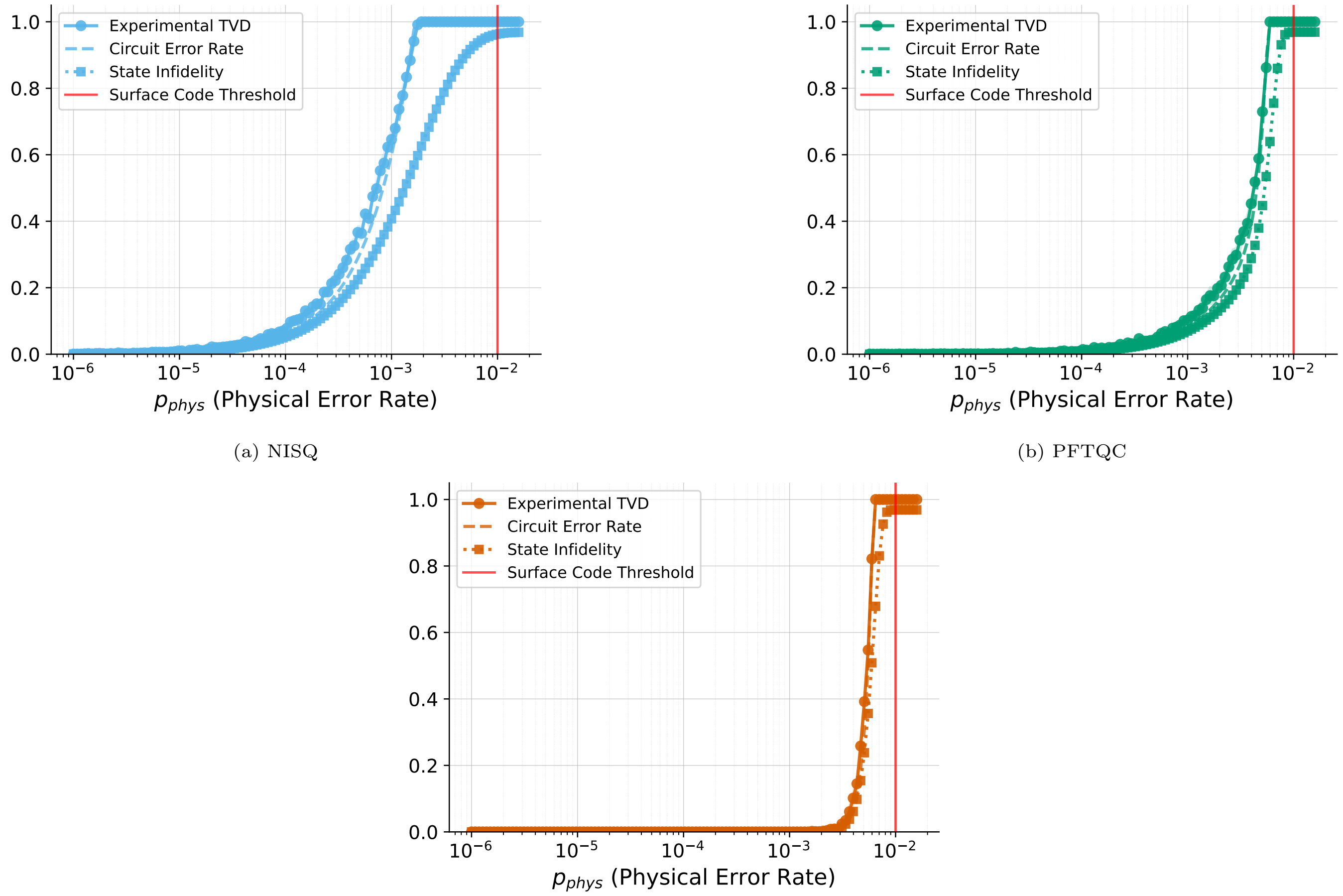
$$\begin{aligned}\mathrm{Tr}(\rho^{(0)}_{out,id}\rho^{(0)}_{noisy}) &= \mathrm{Tr}\Big(\,|\psi\rangle\langle\psi|\sum_j \mu_j\,|\phi_j\rangle\langle\phi_j|\,\Big)\\ &= \sum_j \mu_j|\,\langle\psi|\phi_j\rangle\,|^2\\ &\geq 0,\end{aligned} \quad \text{(L7)}$$

along with the upper-bound previously derived for the total logical circuit error rate.

## Appendix M: Physical Resource Analysis for IQP Advantage

Using our logical accreditation protocol, we apply it to the problem of studying IQP advantage (as described in Section IV). We estimate the physical qubit requirements to keep an IQP circuit's TVD below the quantum advantage threshold. Our analysis in Fig. 19 assumes a physical error rate of $p_{\mathrm{phys}} = 10^{-5}$ for 10 and 20 logical qubit computations. For increasing circuit depths, we find the minimum code distance required by the NISQ, PFTQC, and FTQC regimes to keep the TVD below the quantum advantage threshold. If a regime fails at a given depth, it is considered unviable for deeper circuits. The optimal strategy for an IQP circuit with a given number of layers is then simply the viable regime with the lowest physical qubit cost.

The results show resource-efficiency crossovers between the regimes. NISQ is only optimal for very shallow circuits, after which the PFTQC scheme becomes cheaper. As the circuit deepens further, however, the error from PFTQC's unprotected T-gates forces its required code distance to become prohibitively large. At this point, the FTQC scheme, despite the overhead of magic state distillation, becomes the most efficient strategy.



(a) NISQ

(b) PFTQC

(c) FTQC

FIG. 18: *Comparison of logical accreditation TVD bound and the logical state infidelity for the NISQ, PFTQC and FTQC regimes.* This data is from experiments involving 5-qubit IQP circuits with 40 layers.

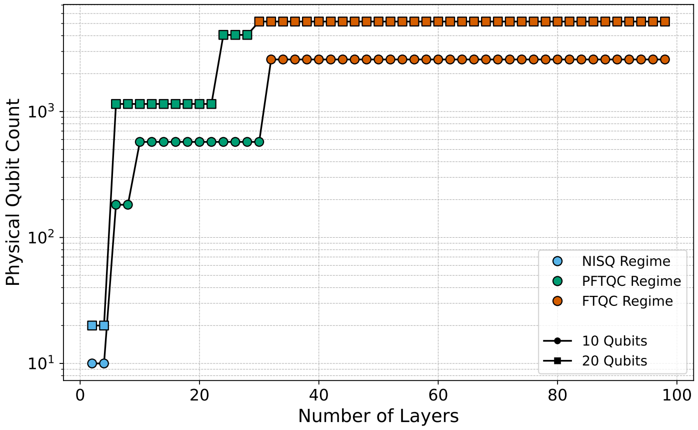


FIG. 19: *An indicative example of the minimum physical qubit count for achieving the IQP threshold for quantum advantage:* This plot shows the minimum required physical qubits as a function of circuit layers for systems with 10 and 20 logical qubits. The color of each point indicates the most resource-efficient regime: NISQ (blue), PFTQC (green), or FTQC (orange). The solid black lines trace the optimal strategy, revealing the crossover points where switching regimes. Here, we take a very optimistic physical error rate of $10^{-5}$. We find the minimum surface code distance (and hence the minimum number of physical qubits) required such that the TVD value is lower than 1/384 (the advantage bound from [41]).