

Parallel Logical Measurements via Quantum Code Surgery

Alexander Cowtan,^{1,*} Zhiyang He², Dominic J. Williamson³, and Theodore J. Yoder⁴

¹Department of Computer Science, *University of Oxford*, Oxford OX1 3QD, United Kingdom

²Department of Mathematics, *Massachusetts Institute of Technology*, Cambridge, Massachusetts 02139, USA

³IBM Quantum, *IBM Almaden Research Center*, San Jose, California 95120, USA

⁴IBM Quantum, *IBM T.J. Watson Research Center*, Yorktown Heights, New York 10598, USA

 (Received 9 August 2025; revised 21 January 2026; accepted 7 April 2026; published 8 May 2026)

Quantum code surgery is a flexible and low-overhead technique for performing logical measurements on quantum error-correcting codes, which generalizes lattice surgery. In this work, we present a code surgery scheme, applicable to any qubit stabilizer low-density parity check (LDPC) code, that fault-tolerantly measures many logical Pauli operators in parallel. For a collection of logically disjoint Pauli product measurements supported on t logical qubits, our scheme uses $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ ancilla qubits, where $\omega \geq d$ is the maximum weight of the single logical Pauli representatives involved in the measurements and d is the code distance. This is all done in time $\mathcal{O}(d)$ independent of t . Our proposed scheme preserves both the LDPC property and the fault-distance of the original code, without requiring ancillary logical codeblocks, which may be costly to prepare. This addresses a shortcoming of several recently introduced surgery schemes, which can only be applied to measure a limited number of logical operators in parallel if they overlap on data qubits.

DOI: [10.1103/gj8x-n5gg](https://doi.org/10.1103/gj8x-n5gg)

I. INTRODUCTION

Topological codes, specifically surface codes, have long been the leading candidates for the practical implementation of large-scale fault-tolerant quantum computation. Recently, experiments have demonstrated subthreshold scaling of surface codes [1,2] and logical operations on 2D color codes [3–5]. Despite their many desirable properties, topological codes in flat space have a notable limitation: their low encoding rate leads to a significant space overhead in achieving fault tolerance. This limitation motivates the study of quantum low-density parity check (LDPC) codes (see Ref. [6] for a recent review), which have higher rate and therefore serve as candidates for low-overhead fault-tolerant quantum memories [7,8].

For topological codes, particularly surface codes [9], methods for logical computation are well-studied [10–12]. Logical computation with quantum LDPC codes, however, is an ongoing research area. Several methods have been proposed, such as constructing codes with native transversal gates [13–25], homomorphic logical

measurements [26,27], and generalized surgery [28]. Transversal gates can in principle be performed in constant time [29,30] and with no additional overhead on the quantum code given a sufficiently powerful decoder, but cannot be universal by the Eastin-Knill theorem [31], and also require a quantum code to have particular structure [32]. Homomorphic measurements and generalized code surgery (also known as LDPC code surgery or simply code surgery) are both techniques that perform logical Pauli measurements on logical qubits, which makes them well-suited to Pauli-based computation [33] where Pauli measurements and magic states together enable universal quantum computation [34]. Nonetheless, homomorphic logical measurements are, thus far, restricted to topological codes and homological product codes. By contrast, generalized code surgery is applicable to any LDPC code [35].

To perform code surgery, we initialize an ancilla system, consisting of new qubits and stabilizer generators, to deform the quantum code in such a way that chosen logical operators are decomposed into stabilizers, a procedure that can be viewed as gauge-fixing or weight reduction [28,36–41]. Code surgery performs a logical measurement on the decomposed operators, which can be made fault-tolerant by measuring all of the new stabilizer generators for a sufficient number of rounds (typically taken to be the code distance d) to prevent timelike errors. After obtaining the logical measurement outcome, we then measure out the ancillary qubits and return to the original code.

*Contact author: akcowtan@gmail.com

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

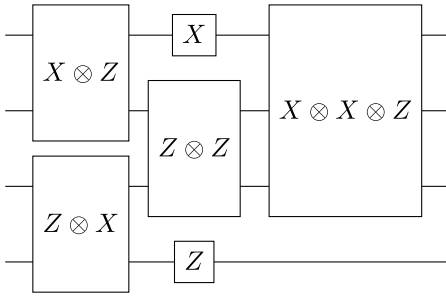


FIG. 1. An example logical circuit with three timesteps of Pauli product measurements. Each timestep contains logically disjoint Pauli products, meaning that they act on different logical qubits, see Definition 1. The logically disjoint Pauli products are measured in parallel.

In this work, we present a code surgery scheme that performs parallel measurements of logical Pauli operators, applicable to any LDPC codes. For a collection of logically disjoint Pauli product measurements (see Definition 1 and Fig. 1) supported on t logical qubits, our scheme uses $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ ancilla qubits, where $\omega \geq d$ is the maximum weight of the single logical Pauli representatives involved in the measurements.

This provides access to a wide array of parallel logical operations, which can be complemented by transversal magic gates or magic state preparation to achieve universal fault-tolerant quantum computation.

A. Related literature

In this work, we build upon substantial recent improvements to quantum LDPC code surgery, which we nonexhaustively review here, eliding many subtleties. Quantum LDPC code surgery was first investigated by Cohen *et al.* [28] using Tanner graphs, with a space overhead of $\mathcal{O}(\omega d) \geq \mathcal{O}(d^2)$ per logical qubit to be measured, given logical operator representatives with weight ω . This space overhead was required to guarantee fault tolerance. We call this the CKBB scheme from now on. Cowtan and Burton [42] took a different approach, using homology, yielding a similar construction where fault tolerance is not guaranteed but the overhead is reduced. This was found to be sufficient to preserve the distance in small practical cases [43]. Shortly after Ref. [43], the independent work of Cross *et al.* [44] showed that if the measured logical operators have suitable expansion properties, then the overhead of the CKBB scheme can be qualitatively reduced without a guarantee that the deformed code is LDPC. Later, Ide *et al.* [41] and Williamson and Yoder [40] independently proposed that one can directly construct an ancilla system with sufficient expansion (instead of relying on the existing expansion properties of the operators) to measure an arbitrary Pauli logical operator with space overhead

$\mathcal{O}(\omega \log^3 \omega)$, where the $\log^3 \omega$ factor comes from graph decongestion [45] and the deformed code remains LDPC.

Remark 1. More precisely, the decongestion overhead was not discussed in Ref. [41], beyond stating that it is polylogarithmic, as they found that in practice it was not required for the codes of interest. In the first arXiv preprint version of Ref. [40] it was erroneously claimed that the decongestion overhead was $\mathcal{O}(\log^2 \omega)$.

Performing Pauli product measurements of disjoint logical operators was further improved by using adapters to connect their ancilla systems in Swaroop *et al.* [46], building on earlier ideas of bridging from Ref. [44] and applying the gauging measurement directly to disjoint logical operators in Ref. [40]. In summary, current methods require $\mathcal{O}(r\omega \log^3 \omega)$ ancillary qubits to measure the product of r logical Pauli operators that are disjoint on the physical data qubits.

Unfortunately, while there have been vast improvements to the original $\mathcal{O}(d\omega)$ overhead of the CKBB scheme, parallelization has presented a persistent obstacle. This obstacle is present because, generically, the logical operators to be measured in a high-rate LDPC code overlap substantially on data qubits as in Fig. 2(b), with the overlap growing with the rate and distance of the code. Consequently, deforming a code without drastically increasing stabilizer weights or stabilizer overlap, and thereby losing the LDPC property, is challenging.

To address the parallelization challenge, Zhang and Li [47] proposed techniques called brute-force branching and devised sticking, such that logical operators of the same type, say $\bar{Z}$ Pauli product measurements, can be measured in parallel using a single ancilla system. At a high level, brute-force branching uses *branching stickers* to create new, disjoint representatives of logical operators on ancilla qubits, which are then measured using the CKBB scheme. Devised sticking, on the other hand, measures all logical operators in a region of the code at once using the CKBB scheme, but requires additional checks and qubits to avoid measuring unwanted logical operators.

Zhang and Li's techniques have three main limitations: (a) they use the CKBB scheme for measurement, which leads to worst-case space overheads of $\mathcal{O}(t\omega(\log t + d))$ and $\mathcal{O}(t^2\omega d)$ for measurement by brute-force branching and devised sticking respectively, both of which are at least quadratic in d as $\omega \geq d$ [48]. (b) Their scheme cannot measure operators containing Y terms in parallel without using ancillary logical $|Y\rangle$ states, as proposed by Ref. [49, Fig. 10] in the context of twist-free lattice surgery. To prepare these logical $|Y\rangle$ states, Zhang and Li required that the LDPC code used must support a transversal logical S gate which is a nontrivial assumption. Distillation, as an alternative, would be much more costly. Moreover, even to measure operators containing a mixture of X and Z

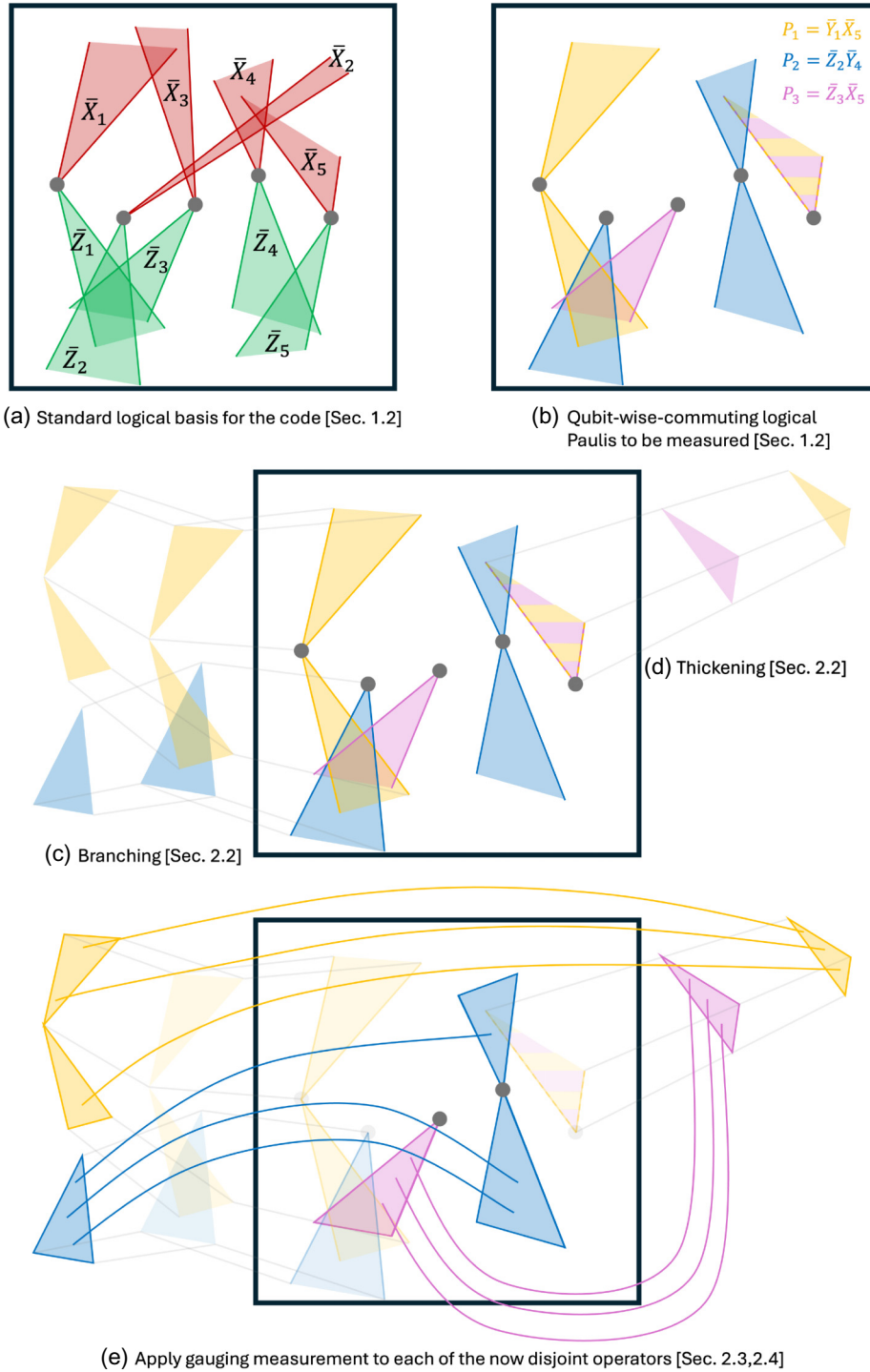


FIG. 2. A summary of our core construction. (a) We start with an arbitrary quantum low-density parity check code and a standard logical basis [50] in which $\bar{X}_i$ and $\bar{Z}_j$ intersect on exactly $\delta_{ij} \in \{0, 1\}$ qubits. (b) Our goal is to measure a set of logical Paulis that, when decomposed in the standard basis, are qubit-wise-commuting, by which we mean that the operators' supports on each logical qubit are either equal or identity. (c),(d) Use low-density parity check surgery techniques, brute-force branching [47] and thickening [40], to create completely disjoint representatives of the logical basis operators that are required. (e) Apply gauging measurement [40], possibly with adapters [46] between disjoint components, to measure the desired set of logical Paulis. Most generally, to instead measure arbitrary commuting sets of logical Paulis, we can use this core construction a constant number of times in sequence, see Sec. III F.

terms in parallel Zhang and Li required ancillary logical $|0\rangle$ states. (c) Their scheme is specific to Calderbank-Shor-Steanne (CSS) codes, rather than more general stabilizer codes.

B. Contributions

Our surgery scheme features two important improvements. First, we use the gauging measurement scheme of Ref. [40] instead of the CKBB scheme, which leads to a significantly lower space overhead and more rigorous guarantees on fault tolerance. Second, we consider the problem of measuring a logically disjoint collection of Pauli product operators, defined as follows.

Definition 1. For a quantum stabilizer code, suppose we fixed a basis of X -type and Z -type logical operators $\bar{X}_1, \dots, \bar{X}_k$ and $\bar{Z}_1, \dots, \bar{Z}_k$ for the k logical qubits, thereby fixing Y -type logical operators $\bar{X}_i \bar{Z}_j$ up to scalars. A logical Pauli product operator P is simply a product of these single logical qubit operators, up to multiplication by stabilizers. The logical support of P is the set of logical qubits on which P acts nontrivially. A collection of logical Pauli product operators $P_1, \dots, P_\ell$ is logically disjoint if their logical supports are disjoint.

Intuitively, logically disjoint Pauli product operators act on different *logical* qubits, i.e., there is no overlap in the logical qubit support. For example, $\bar{X}_1 \otimes \bar{Z}_2$ and $\bar{Z}_3 \otimes \bar{X}_4$, as shown in the first timestep of Fig. 1, are logically disjoint, but $\bar{X}_1 \otimes \bar{Z}_2$ and $\bar{Z}_2 \otimes \bar{Z}_3$, shown in timesteps 1 and 2, respectively, are not. Logically disjoint Pauli product operators need not act on different *physical* qubits, and the physical qubit supports may overlap substantially.

Our scheme can measure logically disjoint mixed type operators that contain logical Y terms in parallel without using ancilla $|Y\rangle$ states or codeblocks. This is a stronger result than the scheme for measuring operators of the same type (X or Z) introduced in Ref. [47]. In the same vein, our scheme can also measure logical operators in parallel on stabilizer codes which are not CSS.

We combine several elements from the recent literature to obtain our results. First, we use a particular basis [50, Sec. 4.1] and a generalized version of brute-force branching from Ref. [47] to obtain copies of the logicals to be measured. Then we construct ancillary expander graphs for each measurement, following Refs. [40,41], and use the adapters from Ref. [46] to connect the ancillary systems to generate product measurements.

At a high level, the overall space overhead of our scheme comes from the fact that brute-force branching t logicals, with weight up to ω each, costs $\mathcal{O}(t\omega \log t)$, and then measurement requires a further $\mathcal{O}(t\omega \log^3 \omega)$. The overall scheme described in this paper is summarized in Fig. 2.

There are two ways to extend our scheme to measure more general collections of Pauli operators.

- (1) While we state our result in terms of logically disjoint products, the same procedure applies to nondisjoint Pauli products that have the same or identity action on every logical operator. For example, we can measure $\bar{X}_1 \bar{Z}_2 \bar{I}_3$ simultaneously with $\bar{I}_1 \bar{Z}_2 \bar{X}_3$, but not $\bar{X}_1 \bar{Z}_2 \bar{I}_3$ with $\bar{Z}_1 \bar{X}_2 \bar{I}_3$. In this slightly more general case, the asymptotic space overhead of $\mathcal{O}(T\omega(\log T + \log^3 \omega))$ is the same except T is now the total number of nonidentity logical Pauli terms in the Pauli products (e.g., $T = 4$ for the above example, the sum of the logical weights of $\bar{X}_1 \bar{Z}_2 \bar{I}_3$ and $\bar{I}_1 \bar{Z}_2 \bar{X}_3$).
- (2) In Sec. III F, we extend our scheme to measure arbitrary commuting subgroups of logical Pauli operators, using twist-free surgery techniques from Refs. [47,49]. These techniques require ancilla logical states, which can be encoded in the same code block or additional code blocks. The ability of our scheme to directly measure logical Y terms in parallel significantly reduces the overhead in preparation of such logical ancillary states as compared to Ref. [47].

In Table I, we compare prior schemes to our own. Column 1 is the CKBB scheme [28] and similar schemes, including Ref. [44]. Column 2 contains the parallel measurement schemes from Ref. [47], which use brute-force branching and devised sticking, respectively. As brute-force branching has a lower asymptotic overhead we use that overhead for the comparison. Column 3 contains the various schemes based on expander graphs, Refs. [40,46,51]. There are many subtleties not included in the table. For example, in columns 1 and 2, Refs. [28,47] require moving to a subsystem code, which can prevent a threshold, while Ref. [44] is not guaranteed to be LDPC, but can obtain improved overheads in practical cases. In column 3, Refs. [40,46] apply to generic stabilizer codes, while Ref. [51] is specified only for CSS codes. Lastly, there are two different definitions of parallelism used in Ref. [47]:

- (1) “Conventional” parallelism, whereby logical operations are applied simultaneously if they act on different logical qubits, see Definition 1.
- (2) “Ultimate” parallelism, whereby logical operations are applied simultaneously if they commute.

In particular, in the latter case the logical operations may overlap on logical qubits. While the majority of this work focusses on the former case, which is reflected in Fig. 1 and Table I, we also extend our results to the latter in Sec. III F. In that extension, the scheme is no longer logical

TABLE I. Comparison of different surgery schemes. The *space cost* is the asymptotic space overhead of a scheme to measure t logical Pauli operators with maximum representative weight ω while maintaining the code distance d . Logarithmic factors are suppressed by $\tilde{\mathcal{O}}(\cdot)$ notation. A scheme is labeled *parallel* when, for a generic low-density parity check code and set of logical operators to be measured, the scheme maintains the low-density parity check property while measuring the logical operators in $\mathcal{O}(1)$ logical timesteps, each taking up to $\mathcal{O}(d)$ rounds of syndrome extraction. A scheme is labeled *S gate-free* when it does not require transversal S gates or $|Y\rangle$ -state distillation. A scheme is *stabilizer* when it is applicable to generic stabilizer codes, rather than just Calderbank-Shor-Steane codes. A scheme is *logical ancilla-free* when the scheme can be applied without requiring any *logical* ancilla qubits on top of the prior *physical* qubit space overhead.

Scheme	Refs. [28,44],	Ref. [47]	Refs. [40,46,51]	This work
Space overhead	$\mathcal{O}(t\omega d)$	$\tilde{\mathcal{O}}(t\omega d)$	$\tilde{\mathcal{O}}(t\omega)$	$\tilde{\mathcal{O}}(t\omega)$
Parallel	$\times$	$\checkmark$	$\times$	$\checkmark$
S gate-free	$\checkmark$	$\times$	$\checkmark$	$\checkmark$
Stabilizer	$\times$	$\times$	$\checkmark$	$\checkmark$
Logical ancilla-free	$\checkmark$	$\times$	$\checkmark$	$\checkmark$

ancilla-free, but all the remaining rows in Table I are identical.

Remark 2 (Weights of logical representatives). Existing methods of LDPC code surgery deform the local code around logical representatives, hence the cost in terms of additional qubits and checks is reliant on the choice of representatives and their weights.

While there are codes for which such Pauli representatives can always be chosen to have weight d , or weights bounded by $\mathcal{O}(d)$ as the blocklength increases, this is not true in general. 3D toric codes, for example, have $\bar{X}$ logical operators, which are asymptotically larger than d [52].

As a consequence, to be fully general, our costs must be computed in terms of ω , the largest logical representative weight of the Paulis to be measured, rather than d . Furthermore, our surgery scheme requires the logical operator representatives $\bar{X}_1, \dots, \bar{X}_k$ and $\bar{Z}_1, \dots, \bar{Z}_k$ to satisfy a set of conditions regarding their supports. While such a basis always exists in any stabilizer code [50], these representatives could have weights that are asymptotically far larger than d , bounded by $\mathcal{O}(n)$ in general. This is unsurprising, as we require no additional structure of the LDPC code which we are performing code surgery on, and good quantum LDPC codes are known to exist [53], hence the logical representative weights to be measured cannot be bounded in general by anything less than $\mathcal{O}(n)$.

C. The setting

First, we assume that we can modify the qubit connectivity between rounds of logical measurements, so we can initialize new ancilla systems with very different connectivity for each round of logical measurements without concern. This is different from the setting of Refs. [44,54,55], in which ancilla systems with fixed connectivity are constructed that can be toggled to measure different serialized Pauli products.

Second, we focus on the asymptotic regime, rather than considering finite-size codes with specific (pseudo-)thresholds. For finite-size codes, we expect that our procedure can be substantially optimized to reduce the overhead far below what the asymptotic upper bound indicates, following Refs. [43,44]. For example, it has been found that a variety of near-term codes do not require decongestion of the expander graphs used for measurement [40,41]. Furthermore, while brute-force branching is asymptotically efficient, the constant factor overheads are formidable and so minimizing the branching cost is vital for our scheme to be efficient on low-blocklength codes. We leave such numerical optimization to future work.

Third, the noise model we consider in our notion of fault tolerance is a phenomenological Pauli and measurement noise model, where Pauli faults may occur on data qubits or measurement outcomes. The phenomenological noise model is justified in Appendix D.

II. PRELIMINARIES

A. Stabilizer codes and Tanner graphs

Here we introduce some basic definitions and lemmas, which are used later.

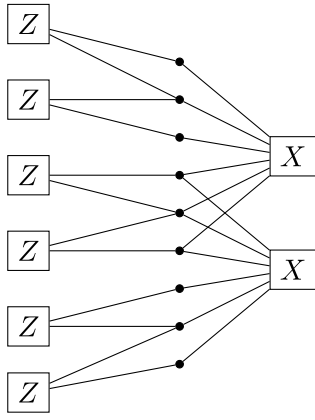
Definition 2. Let $\mathcal{P}^n$ be the Pauli group over n qubits. A qubit stabilizer code Q is specified by an Abelian subgroup $\mathcal{S} \subset \mathcal{P}^n$ such that the codespace $\mathcal{H}$ is the mutual $+1$ eigenspace of $\mathcal{S}$, i.e.,

$$U|\psi\rangle = |\psi\rangle \quad \forall U \in \mathcal{S}, |\psi\rangle \in \mathcal{H}.$$

We say $\mathcal{S}$ is the stabilizer group for Q .

A qubit stabilizer code Q can be specified by a stabilizer check matrix

$$H = [H_X | H_Z] \in \mathbb{F}_2^{r \times 2n}, \text{ such that } H \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} H^\top = 0,$$

FIG. 3. A Tanner graph for the $[[9, 1, 3]]$ Shor code.

where a row $[u|v]$ corresponds to a generator of the stabilizer group, and therefore check on Q , $i^{u \cdot v} X(u)Z(v)$, for $u, v \in \mathbb{F}_2^n$.

Definition 3. A qubit CSS code Q is a qubit stabilizer code where the generators of $\mathcal{S}$ can be split into two sets $\mathcal{S}_X$ and $\mathcal{S}_Z$. $\mathcal{S}_X$ contains Pauli products with terms drawn from $\{X, I\}$ and $\mathcal{S}_Z$ terms drawn from $\{Z, I\}$.

Thus there is a stabilizer check matrix H for Q such that

$$H = \begin{pmatrix} H_X & 0 \\ 0 & H_Z \end{pmatrix}, \quad H_X H_Z^\top = 0.$$

We commonly conflate a Pauli $Z(v)$ operator with its vector v in $\mathbb{F}_2^n$, and the same for a Pauli $X(u)$ operator.

Definition 4. An LDPC code family is an infinite family of stabilizer codes such that the number of qubits participating in each check operator and the number of stabilizer checks that each qubit participates in are both upper-bounded by a constant.

Logical operators in a stabilizer code are Pauli products that commute with all the checks of the code. As the stabilizer group is Abelian, all stabilizers are equivalent to the trivial logical operator. The code distance d of Q is the minimum Pauli weight of all the nontrivial logical operator representatives.

We can use Tanner graphs to describe stabilizer codes, see Fig. 3.

Definition 5. A Tanner graph is a bipartite graph with a vertex for each qubit and each check. A qubit is connected to each check in which it participates with an edge labeled $[1|0]$, $[0|1]$, or $[1|1]$ depending on whether the check acts on the qubit as X , Z , or Y .

In our convention, data qubits are black circles and checks are boxes. If a check only measures Z or X terms

then we omit the edge label, and instead label the box with a Z or X . Therefore for a CSS code we have no edge labels, and all boxes contain either Z or X labels. In this case, the condition that all stabilizers commute is the same as saying that the number of paths between every Z and X check is even.

Depicting large Tanner graphs directly becomes unwieldy, so we use *scalable* notation [56] to make them more compact. Qubits are gathered into disjoint named sets $Q_0, Q_1, Q_2, \dots$, and the same for checks $C_0, C_1, C_2, \dots$. An edge between Q_i and C_j is then labeled with the stabilizer check matrix of Q restricted to Q_i and lifted to C_j , so we have

$$[C_X|C_Z] \in \mathbb{F}_2^{|C_j| \times 2|Q_i|}.$$

If that check matrix is all zeros then omit the edge entirely. If the checks are all of the same type, then we omit the all-zeros half of the check matrix; for example, if C_j contains only Z checks then we label the check box Z and the edge

$$[C_Z] \in \mathbb{F}_2^{|C_j| \times |Q_i|}.$$

For example, in Fig. 4 we show some basic scalable Tanner graphs with qubit sets drawn as circles and check sets drawn as boxes. We have (a) a generic stabilizer code, (b) a CSS code, and (c) the hypergraph product [57] of a classical code C , with bits $\mathcal{L}$ and check matrix ∂_1 , by a repetition code R with blocksize 2 and check matrix $(1 \ 1)$. See Appendix A 1 for a concise summary of hypergraph product codes. We use Tanner graphs, particularly scalable Tanner graphs, extensively to describe our procedures.

Throughout this work, we build on prior literature on quantum LDPC code surgery. We now recount relevant elements that we use.

B. Brute-force branching

Brute-force branching is a method for separating logical operators into new, disjoint supports on the leaves of a tree, see Fig. 5. These new operators are equivalent to the original, but are more convenient for performing logical measurements [47] without losing the LDPC property. After logical measurements, the branches are removed by measuring them out. Branching does not perform any logical operation on the code.

Brute-force branching is formally defined as follows: let Q be a CSS code. Let $\mathcal{I} = \{v_i : i \in 1, 2, \dots, t\}$ be a set of $\bar{Z}$ logicals with support on some data qubits, and let $V = \bigcup_i v_i$ be the union of those supports. Any logical can overlap with any other, and there may be other logicals $\mathcal{J}$ with support wholly in V other than the v_i logicals.

A *brute-force branching* of V is then a tree, which we always take to be a binary tree, of small hypergraph product patches, called branching stickers, connected to V in such a manner as to separate all the logicals v_i into disjoint

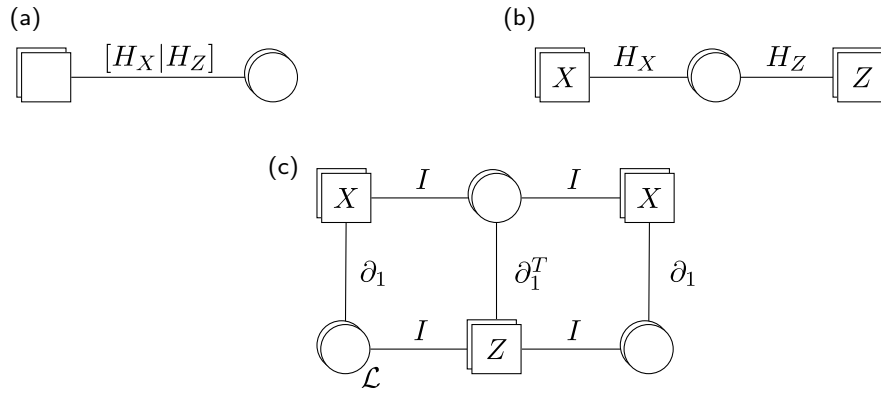


FIG. 4. Examples of scalable Tanner graphs. (a) A generic stabilizer code, (b) a Calderbank-Shor-Steane code, and (c) a hypergraph product code.

supports. V is then the root of the tree, and each leaf contains a copy of each v_i .

Brute-force branching can be described by a scalable Tanner graph, as shown in Fig. 6. Explicitly, let $\mathcal{T}(Q)$ be the Tanner graph of Q . Let $|\mathcal{I}| = t$. Take $\lceil t/2 \rceil$ of the logicals in $\mathcal{I}$ and let the union of their supports on data qubits be S_1 . Let the set of all X -checks incident to S_1 be χ_1 . Including the edges E_1 between S_1 and χ_1 , these generate a bipartite subgraph $G_1(S_1 \cup \chi_1, E_1)$ of the Tanner graph of Q . Then introduce a new graph $G_1^T(S_1^T \cup \chi_1^T, E_1)$. This graph has the same vertices and edges as G_1 , but each vertex in S_1^T is a Z -check, rather than a data qubit, and each vertex in χ_1^T is a data qubit, rather than an X -check. Connect $G_1^T(S_1^T \cup \chi_1^T, E_1)$ to the original Tanner graph of Q by the introduction of edges (u, u') for $u \in S_1$ and its corresponding $u' \in S_1^T$ for each u , and the same for $v \in \chi_1$ and $v' \in \chi_1^T$. Now introduce another new graph H_1 , which is a copy of G_1 . Connect H_1 to G_1^T via edges (u', u) for each $u' \in S_1^T$ and (v', v) for each $v \in \chi_1^T$. Thus far, we have created a single branch for $\lceil t/2 \rceil$ of the logicals in $\mathcal{I}$. The creation of this single branch can be viewed as gluing a hypergraph product code, see Fig. 4(c), to the code Q .

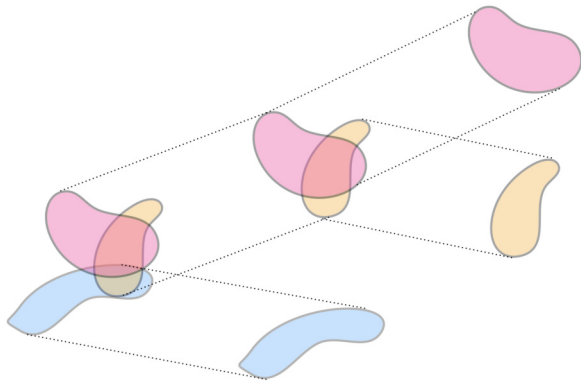


FIG. 5. Illustration of brute-force branching logical representatives with overlapping support.

Repeat this procedure for the other $\lfloor t/2 \rfloor$ logicals: given the union of their supports on data qubits S_2 and χ_2 , all X -checks incident to S_2 , construct $G_2^T(S_2^T \cup \chi_2^T, E_2)$, connect it in a 1-to-1 fashion, then construct H_2 and connect it to G_2^T in a 1-to-1 fashion [58].

Repeat this process $\mathcal{O}(\log t)$ times until we have our t logicals on different stickers on the leaves of the tree. Each leaf corresponds to an equivalent copy of a logical $v_i \in \mathcal{I}$. Hence all t logicals have disjoint support, and we have preserved the LDPC property. If the initial logicals had maximum representative weight ω then each level of the tree uses $\mathcal{O}(t\omega)$ qubits and checks, so the number of additional qubits and checks required for branching will be at most $\mathcal{O}(t\omega \log t)$.

Of course, if there are logicals in V which have no overlap with any other logicals v_i , they do not need to be brute-force branched and can be measured directly. If the initial code Q has a basis of logical operators of both $\bar{Z}$ and $\bar{X}$ type which are disjoint where they commute, as with, e.g., disjoint surface code patches, then brute-force branching is unnecessary.

In Appendix B we demonstrate branching in detail on toy examples.

Lemma 1. Brute-force branching creates a new representative on each leaf for each $v_i \in \mathcal{I}$ and introduces no new logical operators.

Proof. Here we give a proof sketch for readers familiar with the CKBB scheme [28]. In Appendix A2 we introduce a new perspective on brute-force branching as a series of mapping cylinders. This provides an algebraic and formal proof, but which requires some knowledge of elementary homological algebra.

For each $v_i \in \mathcal{I}$, consider the path P down the binary tree from the root to its corresponding leaf. Multiply v_i by the Z checks connected to data qubits in v_i on the first layer, on the branch in the path P . As v_i is a logical operator, this product of Z checks leaves no support in layer 1 and cleans

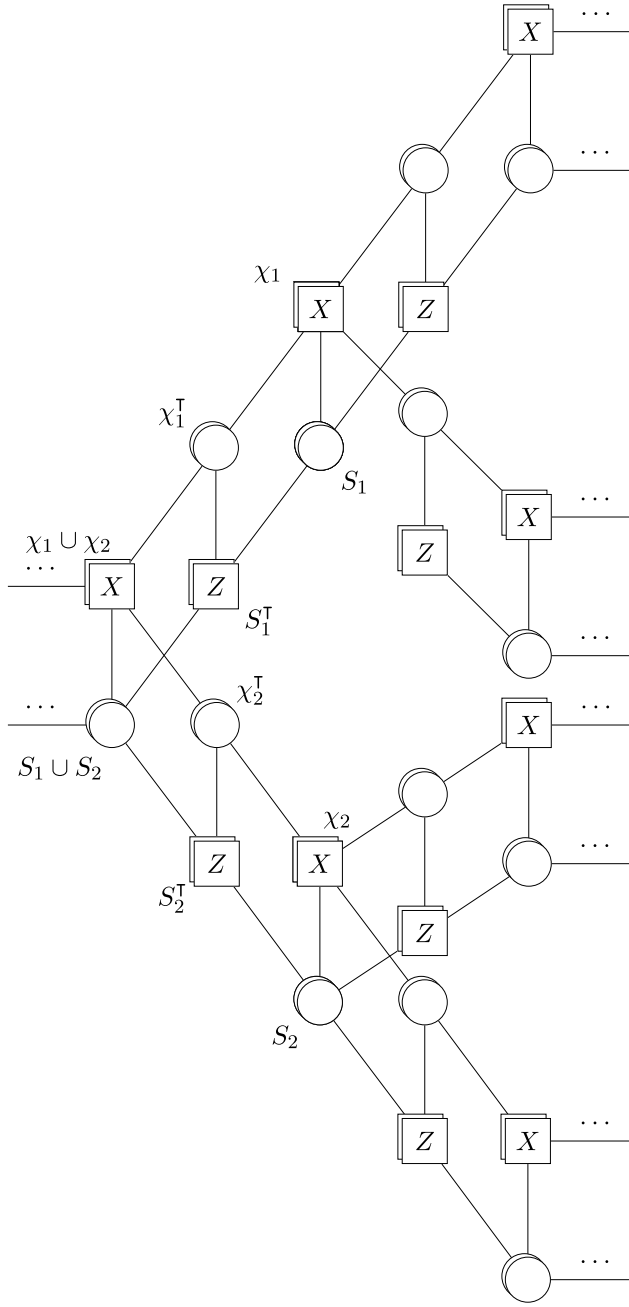


FIG. 6. Scalable Tanner graph for brute-force branching. Starting with a union of logicals V , each layer of the binary tree splits a union of logicals in two, until we are left with disjoint logicals on leaves. In this Tanner graph we do not include the checks and data qubits in the rest of the initial code Q , leaving only dangling edges on the l.h.s. to suggest their existence. Similarly, we leave dangling edges on the r.h.s. to suggest the existence of further branching layers.

v_i to one of the branches in layer 2. Continue this procedure along P , until all support has been moved to the leaf corresponding to v_i . This logical is then another representative of the same operator as v_i .

When measuring $\bar{Z}$ logicals in the CKBB scheme, operators corresponding to new logical qubits (gauge logicals) appear precisely at every odd layer of the attached Tanner graph, that is, Z layers [28]. Let ∂_1 be the local code of Z logicals and incident X checks. Then, in the CKBB scheme, the data qubits on odd layers and their incident Z checks form a code ∂_1^T . As $\ker(\partial_1^T) \neq 0$ in general, these can form new gauge $\bar{X}$ logicals, and indeed these are the only new logicals that can appear, other than their anticommuting $\bar{Z}$ partner logicals [28].

When performing brute-force branching, all these new gauge logicals are actually stabilizers. For any new gauge logical g in an odd layer l , apply the corresponding X checks in layer $l+1$. This creates no new X support on layer $l+1$, as $g \in \ker(\partial_1^T)$ by definition, but does create some X support on layer $l+2$. This new X operator must also not violate any Z checks on this layer, as it is a product of a logical and a stabilizer. Continue applying X checks until the last layer, at which point there are no new layers upon which to add X support, and hence the gauge logical was a product of X checks, and so a stabilizer.

This proof is essentially a cleaning argument, cf. Ref. [59, Lemma 1]. ■

Remark 3. As a consequence of Lemma 1, unlike in the CKBB measurement scheme, we do not have to move to a subsystem code when brute-force branching. The CKBB measurement scheme for a $\bar{Z}$ operator ends with a layer of Z -checks, rather than a layer of X -checks when brute-force branching, which leads to gauge operators which are inequivalent to stabilizers.

Lemma 2 ([47, Theorem 2]). Brute-force branching cannot reduce the code distance.

Proof. Assume we are branching $\bar{Z}$ logicals w.l.o.g. on a starting code Q with distance d . As there are no new logical qubits, to prove that the code distance is preserved we only need to show that multiplying old logicals in the starting code by stabilizers in the new code does not reduce the distance.

For the X distance, observe that all new X checks are on new data qubits, but there are old X checks which have been deformed to have support on the first layer of the branching. Applying these old X checks cannot reduce the weight of an initial $\bar{X}$ operator's support on Q below d , as it only has new support on new qubits. In other words, if we take an operator $\bar{\Lambda}_X$ on Q and apply deformed checks C_D , the restriction of $C_D \bar{\Lambda}_X$ to Q must have weight at least d . The new X checks only have support on new qubits, and so cannot reduce the weight of any operator on Q . Therefore the X distance is at least d .

For the Z distance, we start with an arbitrary operator $\bar{\Lambda}_Z$ on Q . Applying a new Z check in the first layer moves the support of $\bar{\Lambda}_Z$ from support of the check in Q to their

matching data qubits in the second layer of that branch, which is an X layer, potentially leaving behind more support in Q and on the first layer. We can repeat this process, but there are always X layers after every Z layer so we always add more support, and so the weight of $\bar{\Lambda}_Z$ cannot be decreased, hence the Z distance is at least d . ■

Remark 4. The reason why the CKBB scheme [28] requires $\mathcal{O}(d)$ layers to preserve the distance, rather than only two layers for each branch when brute-force branching, is also because in the CKBB measurement scheme for a Z operator, we end with a Z layer rather than an X layer.

To perform brute-force branching of $\bar{Z}$ logicals, we initialize all new qubits in the $|+\rangle$ basis at the same time and then measure all checks. For $\bar{X}$ logicals, the scheme is dualized and we initialize in the $|0\rangle$ basis instead.

Lemma 3. Brute-force branching is a fault-tolerant operation with fault-distance d , assuming we use at least d rounds of stabilizer measurements. Unbranching by measuring out ancilla qubits and measuring all stabilizers in the original code is also a fault-tolerant operation, but only requires one round of stabilizer measurements.

Proof. See Appendix D 1. ■

Despite brute-force branching originally being defined for X or Z -type operators on a CSS code, it is possible on any stabilizer code under a suitable condition given below. This observation is similar to the approach in Ref. [40] for measurement, but for branching instead.

Lemma 4. Let Q be a stabilizer code. Let $\mathcal{I} = \{\bar{\Lambda}_1, \bar{\Lambda}_2, \dots, \bar{\Lambda}_t\}$ be a set of t logical Pauli representatives such that for each $i, j \in [t]$ and data qubit α in Q either (a) $\bar{\Lambda}_i$ and $\bar{\Lambda}_j$ act on α with the same Pauli operator or (b) at least one of $\bar{\Lambda}_i$ and $\bar{\Lambda}_j$ act with identity on α . Then all t logical representatives can be simultaneously branched onto leaves.

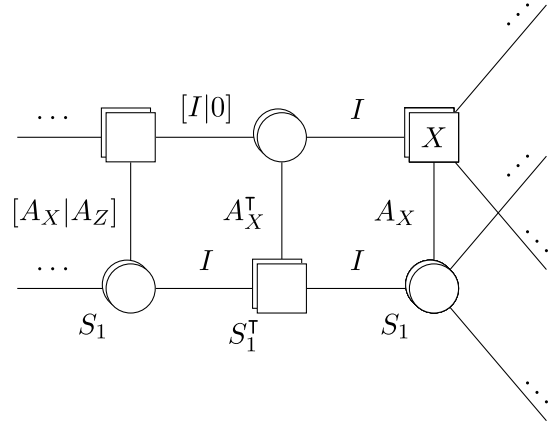
Proof. For each data qubit α in $V = \bigcup_i \bar{\Lambda}_i$, apply a suitable single-qubit Clifford C_α to convert the logical representative into a Z Pauli on that qubit. This permutes the bases of incident checks accordingly. As the logical representatives act with the same Pauli, or the identity, this simultaneously changes the basis of every logical representative in $\mathcal{I}$ on data qubit α to Z or I .

Now, every check on the code which locally anticommutes with the action of any $\bar{\Lambda}_i$ on a given qubit α has a Y_α or X_α term. Considering only the X_α component of a Y_α term, construct the branch accordingly as for $\bar{Z}$ logicals in a CSS code. Last, convert each data qubit back to the original basis by applying $C_\alpha^\dagger$.

As the checks all commute in the Z basis, they must also commute when permuted back to their original basis. Because there exists a product of checks which cleans each logical $\bar{\Lambda}_i$ to a leaf in the Z basis, the same product of checks must clean that logical to a leaf in the original basis. ■

The basis changing is a step performed offline to compute the checks and connectivity required for the branches; the local Cliffords are not intended to be physically applied on a quantum computer.

We illustrate the branch for $S_1 \subset V$ once the basis has been changed to the Z basis below:



$[A_X|A_Z]$ is the check matrix of the code Q restricted to those checks with X -support in S_1 . Dangling edges are left on the l.h.s. to indicate connections to the rest of the code and on the r.h.s. to indicate connections to the rest of the branch. There is a similar, overlapping branch for $S_2 \subset V$. Observe that this S_1 branch is identical to $\bar{Z}$ logical branching in a CSS code, albeit some incident checks to S_1 in the original code have some Z check component, because they involve measurements in the Y basis. This Z check component on the original code does not affect the commutation relations or multiplication of logical operators through to leaves by Z checks in odd layers.

Because this more general branching is equivalent to branching $\bar{Z}$ logicals up to local Cliffords and an irrelevant A_Z check component on the original code, all previous lemmas from this section apply. Another consequence of this generalized branching is that each logical representative on a leaf will be a product of Z Paulis, regardless of the basis it acted on in the original code—whether it is a $\bar{Z}$, $\bar{X}$, or $\bar{Y}$ operator.

C. Gauging logical measurements

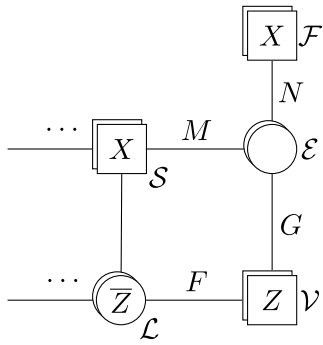
Our procedure for fault-tolerant quantum computation also involves gauging logical measurements [40]. Gauging logical measurements can be applied to any stabilizer code. In the case where a code is CSS the gauging logical

measurement framework coincides with the homological measurement framework of Ref. [41].

Definition 6 ([40]). Let $\mathcal{L}$ be a $\bar{Z}$ logical operator in a stabilizer code. A measurement of $\mathcal{L}$ in the *gauging logical measurement* framework uses a set of appended ancillary qubits and checks forming a graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F})$.

$\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F})$ has a set of vertices $\mathcal{V}$, which are identified with Z checks, in 1-to-1 correspondence with data qubits of $\mathcal{L}$. Each vertex has support on that data qubit in $\mathcal{L}$ given by the correspondence. $\mathcal{G}$ also has a set of edges $\mathcal{E}$, which are identified with data qubits, and some of these are included in deformed X checks which are incident to $\mathcal{L}$. Lastly, $\mathcal{G}$ has a set of faces $\mathcal{F}$, which form a cycle basis of $\mathcal{G}$. These faces are for gauge-fixing, are only connected to edges in $\mathcal{E}$, and do not interact with the original code.

Using a scalable Tanner graph, we have



where $\mathcal{L}$ is the original logical operator and S is the set of all X -checks incident to $\mathcal{L}$. The sets $\mathcal{V}$, $\mathcal{E}$, and $\mathcal{F}$ are newly introduced to perform the measurement, and

- (1) F is a matrix defining the connectivity from Z checks to data qubits in $\mathcal{L}$; for our purposes, $F = I$.
- (2) G is the incidence matrix of the graph $\mathcal{G}$.
- (3) M defines the connectivity from X checks incident to $\mathcal{L}$ to new data qubits. M is determined by the choice of graph and initial logical operator.
- (4) N picks out a cycle basis of $\mathcal{G}$, such that $\ker G = \text{im} N^\top$.

To perform the measurement, initialize all new edges in $\mathcal{G}$ in the $|+\rangle$ basis and measure all stabilizers of the deformed code.

The auxiliary graph $\mathcal{G}$ must satisfy some desiderata to yield a favorable code surgery scheme.

Theorem 1 ([40, 46]). To ensure the deformed code has exactly one less logical qubit than the original code and measures the target logical operator $\bar{Z} = Z(\mathcal{L})$, it is sufficient that

- (1) $\mathcal{G}$ is connected.

To ensure the deformed code is LDPC, it is necessary and sufficient that

- (1) $\mathcal{G}$ has $\mathcal{O}(1)$ vertex degree.
- (2) Each stabilizer of the original code has a short perfect matching in $\mathcal{G}$, and each edge is in $\mathcal{O}(1)$ matchings.
- (3) There is a cycle basis of $\mathcal{G}$ in which (a) each cycle is length $\mathcal{O}(1)$ and (b) each edge is in $\mathcal{O}(1)$ cycles.

To ensure the deformed code has code distance at least the distance d of the original code, it is sufficient that

- (1) $\mathcal{G}$ has relative Cheeger constant $\beta_d(\mathcal{G}, f(\mathcal{L})) \geq 1$.

We elide the details of the expansion properties and perfect matchings, as they are not the focus of this work. It is sufficient for our purposes to say that, given any logical operator, these desiderata can always be satisfied, and in fact Ref. [40] proves that it is easy to do so.

The space overhead of such a graph for measuring any arbitrary logical operator with weight $\mathcal{O}(\omega)$ is $\mathcal{O}(\omega \log^3 \omega)$. The $\log^3 \omega$ factor comes from decongestion [45], whereby the graph is thickened in order to find a low-weight cycle basis, a procedure which can be performed efficiently by a randomized algorithm. By thickening, we add “dummy vertices” which are not connected to the original code Q but are required to maintain the LDPC property.

Remark 5. The desiderata above ensure that the deformed code is LDPC when measuring only one target logical operator or t disjoint logical operators. If there are many representatives to be measured, and they overlap substantially, then Theorem 1 is not sufficient to preserve the LDPC property, as data qubits at the overlap are involved in $\mathcal{O}(t)$ incident checks.

Lemma 5 ([40, Lemma 7]). If all the desiderata of Theorem 1 are satisfied, the gauging measurement procedure given by initializing all new data qubits in suitable bases and measuring all stabilizers for at least d rounds of measurement has fault-distance d .

D. Universal adapters

The last definition we need is that of adapters, whereby ancilla systems for measuring disjoint logical Paulis using the gauging logical measurements formalism can be connected to measure their product.

Definition 7 ([46]). Consider two disjoint logical operators with accompanying graphs $\mathcal{G}_l$ and $\mathcal{G}_r$, defining gauging logical measurement systems for each individual operator. Then a universal adapter $\mathcal{A}$ is a set of additional vertices, edges and faces connected to $\mathcal{G}_l$ and $\mathcal{G}_r$. This converts $\mathcal{G}_l$ and $\mathcal{G}_r$ into a larger, *adapted* graph $\mathcal{G}_l \sim_{\mathcal{A}} \mathcal{G}_r$.

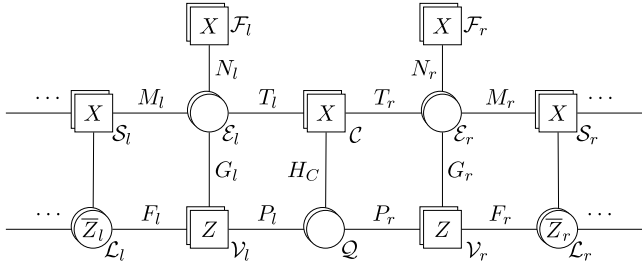


FIG. 7. Measuring $\bar{Z}_l \bar{Z}_r$ without measuring either logical operator individually. The adapter edges joining auxiliary graphs G_l and G_r into one larger auxiliary graph are those hosting the qubits Q . The matrices T_l, T_r, P_l, P_r are efficiently determined using the SkipTree algorithm, [46, Theorem 7]. We do not include in this diagram qubits or checks outside the supports of the logicals $\bar{Z}_l$ and $\bar{Z}_r$, and instead suggest their existence through dangling edges.

Theorem 2 ([46, Theorem 11]). Provided auxiliary graphs to perform gauging measurements of nonoverlapping logical operators $\bar{Z}_0, \bar{Z}_1, \dots, \bar{Z}_{t-1}$ that satisfy the graph desiderata of Theorem 1, there exists an auxiliary graph to perform gauging measurement of the product $\bar{Z}_0 \otimes \bar{Z}_1 \otimes \dots \otimes \bar{Z}_{t-1}$ satisfying the desiderata of Theorem 1 [60].

This gauging measurement first takes the set of operators $\bar{Z}_0, \bar{Z}_1, \dots, \bar{Z}_{t-1}$ and applies adapters as in Fig. 7 between $t-1$ overlapping pairs of disjoint logicals, for example, $\bar{Z}_0 \bar{Z}_1$ then $\bar{Z}_1 \bar{Z}_2$, etc. until all t logicals are connected. In particular, the measurement can be performed while maintaining the LDPC property and fault-distance of the code.

This means that so long as we can separate logical operators to have disjoint support, we can perform Pauli product measurements of those operators freely. Importantly, the adapters add no further asymptotic overhead beyond that of gauging logical measurements, so if we have t Pauli terms in a logical product measurement we can measure products of them with $\mathcal{O}(t\omega \log^3 \omega)$ space overhead.

Remark 6. The gauging measurement can be applied directly to a product of disjoint logical operators, but in the worst case this achieves a less efficient scheme than using adapters. For a product of t terms with maximum representative weight ω , direct gauging logical measurements uses $\mathcal{O}(t\omega \log^3 t\omega)$ additional qubits, while employing adapters brings the qubit cost to $\mathcal{O}(t\omega \log^3 \omega)$. Although the adapter scheme can also be viewed as a large gauging measurement, it structures the creation of the gauging measurement graph in such a way to reduce the space overhead.

Unlike brute-force branching, we do not make use of the finer details of gauging logical measurements or universal

adapters, and they can be considered black-box procedures for the purpose of our work. See Ref. [54] for a high-level overview of the construction of gauging logical measurements and adapters.

III. THE PARALLEL MEASUREMENT PROCEDURE

In this section we first consider only measuring single logical qubits of the same type (X or Z) in parallel on CSS codes. We then incrementally generalize this result to measurements in different bases including Y . In Sec. III D we generalize to stabilizer codes. Then in Sec. III E we show how to measure Pauli products in parallel, and in Sec. III F we show how to measure arbitrary commuting sets of Pauli products.

Definition 8 (Individual Pauli operator). Let Q be a quantum LDPC code, and let $\mathcal{B} = (\mathbb{C}^2)^{\otimes k}$ be a choice of tensor decomposition of the logical space L of Q into qubits, i.e., $\mathcal{B} \cong L$. Then an individual Pauli operator acts on a single copy of $\mathbb{C}^2$ in the tensor decomposition.

In other words, an individual Pauli operator $\bar{P}_i$ acts on only one logical qubit of Q . Clearly, this notion is basis-dependent: if a different basis is chosen, the same physical operator on the code could act on multiple (or different) logical qubits.

Definition 9 (Pauli product). Let Q be a quantum LDPC code, and let $\mathcal{B} = (\mathbb{C}^2)^{\otimes k}$ be a choice of tensor decomposition of the logical space L of Q into qubits, i.e., $\mathcal{B} \cong L$. Then a Pauli product may act on multiple copies of $\mathbb{C}^2$ in the tensor decomposition.

A Pauli product can then be expressed as $\bigotimes_{i \in [k]} \bar{P}_i$ for individual Pauli operators $\bar{P}_i$. Each $\bar{P}_i$ in a Pauli product is called a *Pauli term*, or just *term*, if it is clear from context.

A. Individual logicals of the same type

We first consider the simple case where we have many logical qubits in a CSS code to measure simultaneously in the *same* basis, either Z or X . For example, measuring $\bar{Z}_1$ and $\bar{Z}_2$ simultaneously. We call these *Z-type measurements* and similarly for *X-type measurements*.

The method first uses brute-force branching, then unlike in Ref. [47] the measurements on leaves are performed using expander graphs [40] rather than hypergraph products [28]. We illustrate this procedure in Fig. 8.

Lemma 6. Let Q be a CSS quantum LDPC code. We can simultaneously measure individual Pauli operators of the same type (Z or X), with representative weight at most ω , on t different logical qubits of Q while retaining fault-distance d and the LDPC property, using $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ new data qubits and checks.

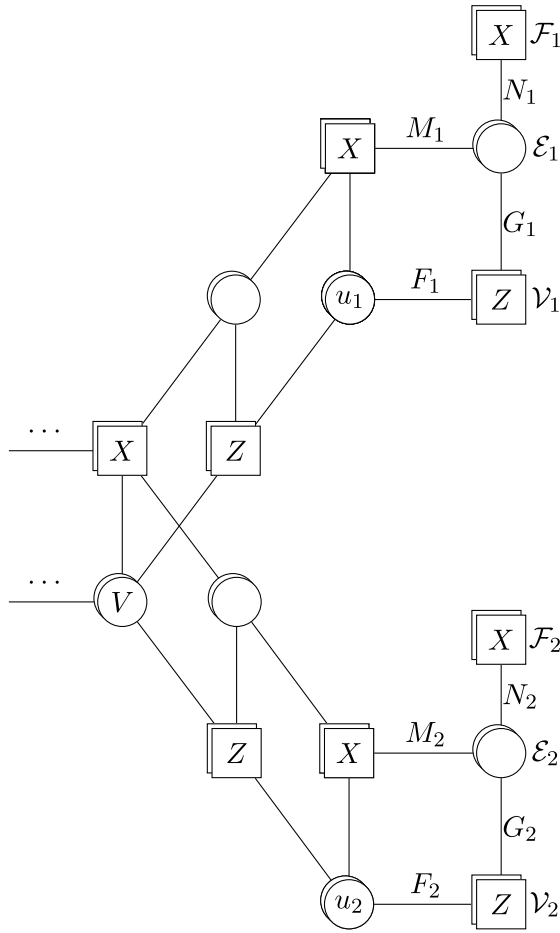


FIG. 8. Branching and gauging measurement for two $\bar{Z}$ logicals. Note that for just two logicals, branching is unnecessary and the gauging measurement systems can be appended in-place, but this scalable Tanner graph generalizes to an arbitrary number of overlapping $\bar{Z}$ logicals.

Proof. Perform brute-force branching on the logicals $v_i \in \mathcal{I}$, using the union V of support on data qubits. Next, we use gauging logical measurements, appending expander graphs to the leaves of the brute-force branching tree, measuring each of the individual logicals we desire to measure. Any logicals which we do not wish to measure, but that are contained in V , are not fully branched, so are left without a leaf or appended expander graph.

By a combination of Lemmas 3 and 5, this procedure is fault-tolerant with fault-distance d , assuming that the brute-force branching step and measurement step are each performed for d rounds of stabilizer measurements.

The brute-force branching process requires $\mathcal{O}(tw \log t)$ additional qubits, and the gauging logical measurement a further $\mathcal{O}(tw \log^3 \omega)$, resulting in a scaling $\mathcal{O}(tw(\log t + \log^3 \omega))$. ■

We next consider measuring single logical qubits in parallel, but where the measurement bases are a mixture of X s

and Z s. For this we need to branch both $\bar{X}$ and $\bar{Z}$ operators at the same time.

B. Individual logicals of CSS type

Lemma 7. Let Q be a CSS quantum LDPC code. Let the unions of $\bar{Z}$ and $\bar{X}$ logical representatives to be measured have no overlap, i.e., no $\bar{Z}$ representative to be measured intersects with any $\bar{X}$ representative to be measured. We can simultaneously measure individual Pauli operators with representative weight at most ω of either Z or X type on t different logical qubits on Q while retaining fault-distance and the LDPC property, using $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ new data qubits and checks.

Proof. As each $\bar{Z}$ logical has no intersection with an $\bar{X}$ logical, each union of $\bar{Z}$ and $\bar{X}$ logicals can be branched independently, and their representative copies on leaves measured by gauging logical measurement. Formally, we can apply local Cliffords to each $\bar{X}$ logical to convert it into a product of Z Paulis and compute the branching of all t logicals as though they are $\bar{Z}$ logicals using Lemma 4, then apply local Cliffords to convert back, to acquire a valid branching for the logicals of mixed types. Hence we can apply Lemma 6. ■

Unfortunately, representatives of $\bar{Z}$ and $\bar{X}$ operators will generally overlap substantially, even when they commute. When the representatives overlap, the new checks introduced in Z -type and X -type branches do not commute in general, as the $\bar{X}$ logicals are deformed when branching $\bar{Z}$ logicals and vice versa [47, Appendix F]. Thus Lemma 7 requires the representatives to be chosen such that no $\bar{Z}$ representative to be measured overlaps with any $\bar{X}$ representative to be measured. This is always possible when the logical operators are on different logical qubits, as stated by the following lemma.

Lemma 8. Consider a CSS code Q with check matrices H_X, H_Z and parameters $[[n, k, d]]$. We can find $\bar{X}$ logical operators $x_1, \dots, x_k \in \mathbb{F}_2^n$ and $\bar{Z}$ logical operators $z_1, \dots, z_k \in \mathbb{F}_2^n$ of this code, such that $|x_i \cap z_j| = \delta_{ij}$. Here $\cap$ denotes intersection as sets.

Corollary 1. In this basis, any set of $\bar{X}$ and $\bar{Z}$ operators acting on different logical qubits have representatives such that their unions have no data qubit overlap.

Remark 7. We consider this lemma a standard result for CSS codes and include a proof in Appendix C for completeness. We are aware of independent proofs, see Refs. [61, Sec. 4] and [62, Sec. 6]. It is also a special case of the “standard form” for stabilizer codes given in Ref. [50, Sec. 4.1].

Remark 8. This choice of basis may give logical operators with substantially greater representative weight than

the minimal weight possible for each logical operator. Lemma 8 gives an upper bound on the representative weight of a $\bar{Z}$ logical of $n - m - k + 1$, so $\omega \in \mathcal{O}(n)$ for LDPC codes. Similarly, the cleaned $\bar{X}$ logical representatives may have larger weight than otherwise. Thus we emphasize that here ω is the largest weight of these chosen representatives, not the minimum available for each logical operator.

Of course, while Lemma 8 shows that there exists a basis such that the unions of $\bar{Z}$ and $\bar{X}$ representatives do not overlap, there may be *many* such bases, and many valid choices of representatives with different weights; consider for example a surface code patch, which has only one logical qubit and many different representatives with different weights satisfying Lemma 8. Thus in practice it will be vital to determine low-weight representatives satisfying the no-overlap condition.

Remark 9. For general quantum LDPC codes, finding a logical basis which possesses minimum-weight representatives is NP-hard [63], and hard to approximate [64], even without stipulating any further conditions such as the no-overlap condition above. Therefore, with general surgery schemes for logical measurement, even where the logical basis does not have conditions on its properties, one must be careful and compute the overheads in terms of ω , rather than assuming each representative has weight d .

As an alternative option, rather than choosing new logical operators that satisfy the no-overlap condition, we could instead start with representatives that do overlap. Next, branch the $\bar{Z}$ logicals and track the overlapping $\bar{X}$ logicals, which are deformed by the branching. Then branch these $\bar{X}$ logicals on top of the deformed code, so there will be some connectivity between the $\bar{X}$ and $\bar{Z}$ branches. When the $\bar{X}$ logicals commute with every $\bar{Z}$ logical contained in $V = \bigcup_i v_i$, the deformed $\bar{X}$ logicals will pick up support in the branches with weight proportional to the number of checks required to clean them from V . When they do not commute with every contained $\bar{Z}$ logical, which can happen if the logical operator basis is not the chosen basis given by Lemma 8, then they may pick up support on up to $\mathcal{O}(t\omega \log t)$ new data qubits in the branches. As $\mathcal{O}(t\omega \log t)$ is not bounded by $\mathcal{O}(n)$ in general, this can be far worse than choosing a suitable basis, but could be preferable at finite sizes. We discuss this further in Appendix E.

C. Individual logicals of non-CSS type

We now consider the case when there are also $\bar{Y}$ Paulis to be measured. Unlike in Ref. [47] we do not use ancilla logical qubits or additional codeblocks for this procedure. Instead, we use the same chosen basis as in Sec. III B.

Corollary 2. Let $\bar{\Lambda}_i$ and $\bar{\Lambda}_j$ be any two arbitrary Pauli operators— $\bar{X}$, $\bar{Y}$ or $\bar{Z}$ —on different logical qubits i, j with representatives given by Lemma 8. Let α be an arbitrary data qubit. Either (a) $\bar{\Lambda}_i$ and $\bar{\Lambda}_j$ act on α with the same Pauli operator or (b) at least one of $\bar{\Lambda}_i$ and $\bar{\Lambda}_j$ act with identity on α .

As a consequence, given any set of $t \leq k$ logical Pauli operator representatives $\bar{\Lambda}_1, \bar{\Lambda}_2, \dots, \bar{\Lambda}_t$ from Lemma 8 on t different logical qubits, these can all be branched such that an equivalent logical operator appears on each leaf.

Lemma 9. Let Q be a CSS quantum LDPC code, with a chosen basis of logical operators given by Lemma 8. We can simultaneously measure any Pauli operators on t different logical qubits individually on Q while retaining fault-distance and the LDPC property. Letting ω be the maximum representative weight of logicals to be measured, this uses $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ new data qubits and checks.

Proof. First branch using Lemma 4. After branching, each representative is a product of Z Paulis on each leaf of the tree. Then use gauging logical measurements to measure each Pauli term on each leaf. By a combination of Lemmas 3 and 5, this operation is fault-tolerant with fault-distance d , given that the branching and measurement steps were each performed with at least d rounds of stabilizer measurements. ■

D. Individual logicals on stabilizer codes

We now show that our protocol for measuring individual logical operators in parallel generalizes to stabilizer LDPC codes which are not CSS codes. For this, we use the “standard form” of stabilizer codes given in Ref. [50, Sec. 4.1]. Recall that in standard form, which can be obtained by Gaussian elimination, the stabilizer generator matrix of an arbitrary stabilizer code Q has the form

$$H = [H_X | H_Z] = \left[\begin{array}{ccc|ccc} I & A_1 & A_2 & B & C_1 & C_2 \\ 0 & 0 & 0 & D & I & E \end{array} \right],$$

where the first three columns are of sizes r , $n - k - r$, k , respectively, and so are the second three columns; r is the rank of H_X as a matrix over $\mathbb{F}_2$. The two rows are of sizes r and $n - k - r$, respectively.

It is then a standard result that we can pick a basis of logical $\bar{X}$ representatives given by the matrix,

$$\Xi_{\bar{X}} = \left[\begin{array}{ccc|ccc} 0 & E^\top & I & V_1 & 0 & 0 \end{array} \right],$$

where each $\bar{X}$ logical is a row x_i in the matrix, and $V_1 = E^\top C_1^\top + C_2^\top$. Any two rows x_i, x_j act in the same basis on any overlapping data qubits. This determines a basis of

$\bar{Z}$ logical representatives, which up to stabilizers can be described by rows in the matrix,

$$\Xi_{\bar{Z}} = \begin{bmatrix} 0 & 0 & 0 & | & A_2^T & 0 & I \end{bmatrix}.$$

Similarly, any two rows z_i, z_j act in the same basis (the Z basis in this case) on any overlapping data qubits. Further observe that any two logicals x_i and z_j where $i \neq j$ overlap only on up to r data qubits, in which they act in the same basis. Otherwise, if $i = j$ they anticommute on exactly one data qubit. Hence Corollary 1 also applies to this basis, and we acquire the following lemma.

Lemma 10. Let Q be a stabilizer quantum LDPC code, with a chosen basis of logical operators given by the “standard form.” We can simultaneously measure any individual Pauli operators on t different logical qubits on Q while retaining fault-distance and the LDPC property. Letting ω be the maximum representative weight of logicals to be measured, this uses $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ new data qubits and checks.

We demonstrate the scheme for parallel measurement on stabilizer codes in detail in Appendix B 2, using the $[[8, 3, 3]]$ code.

E. Measuring Pauli products

Measuring logical Pauli products in parallel is a straightforward extension of the scheme outlined above to measure individual logical Paulis in parallel. The extension is straightforward because we can separate all logicals included in a set of parallel measurements onto disjoint supports by brute-force branching, and then applying universal adapters to the graphs introduced by the gauging measurement leaves Tanner graph connectivity unaffected elsewhere in the code, see Fig. 9.

Theorem 3. Let Q be a stabilizer quantum LDPC code, with a chosen basis of logical operators given by Lemma 8.

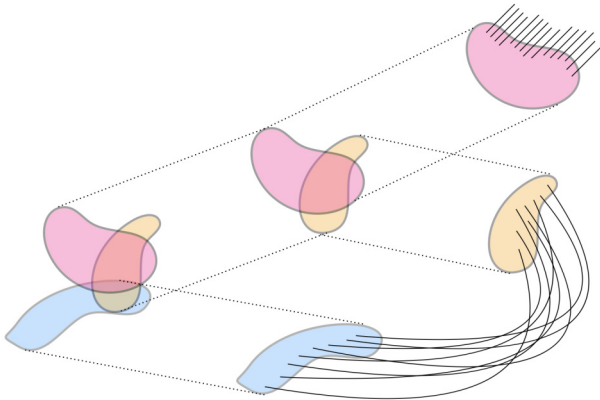


FIG. 9. Illustration of brute-force branching followed by gauging logical measurements, with adapters for Pauli products.

We can simultaneously measure any logically disjoint Pauli products on t logical qubits on Q while retaining fault-distance and the LDPC property. Letting ω be the maximum representative weight of logical terms to be measured, this uses $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ new data qubits and checks.

Proof. To measure Pauli products, we first repeat the procedure from Lemma 10, branching each logical operator term in each product and then performing gauging logical measurements. Now, we also introduce adapters between the measurement graphs for operators whose products are being measured. By Theorem 2 constructing adapters to make a single Pauli product preserves the desiderata of Theorem 1, and so fault-distance and the LDPC property. As each logical appears only once in any given product to be measured, by the assumption that we are measuring t terms on different logical qubits, there are at most two adapters connected to any measurement graph, so the LDPC property is preserved for arbitrary sized products. Similarly, each of the Pauli product measurements acts on disjoint sets of logicals, so the fault-distance of Lemma 5 carries through. The addition of universal adapters does not change the asymptotic space overhead, so the result follows. ■

Note that ω is the largest representative weight of a given Pauli term ($\bar{X}$, $\bar{Y}$, or $\bar{Z}$) included in the measurement, not the largest representative weight of a measured Pauli product.

F. Arbitrary commuting products

The same procedure for measuring logical Pauli products in parallel described above can be extended straightforwardly to the case where we measure logical Pauli products containing terms which have the same, or identity, action on every logical operator.

In this case, the procedure is the same as for Theorem 3, except that when branching individual representatives to leaves we extend the branches [as depicted in Fig. 2(d)] to give multiple copies of each representative on different leaves. For example, if we wish to measure $\bar{X}_1 \bar{Z}_2 \bar{X}_3$ simultaneously with $\bar{I}_1 \bar{Z}_2 \bar{X}_3$, then we have four leaves in total: one for $\bar{X}_1$, two for $\bar{Z}_2$, and one for $\bar{X}_3$. Therefore the space overhead of $\mathcal{O}(T\omega(\log T + \log^3 \omega))$ from Theorem 3 carries through, but T is now the number of nonidentity terms in the set to be measured.

To measure an arbitrary commuting set Θ of logical Pauli products in time independent of t , we cannot branch, say, $\bar{X}_1$ at the same time as $\bar{Z}_1$ in order to measure $\bar{X}_1 \bar{Z}_2 \bar{I}_3$ simultaneously with $\bar{Z}_1 \bar{X}_2 \bar{I}_3$. Therefore we instead use the technique of twist-free surgery from Ref. [49] to separate each Pauli product into a pair of commuting products which do not contain $\bar{Y}$ terms, but do have additional support on logical ancilla qubits.

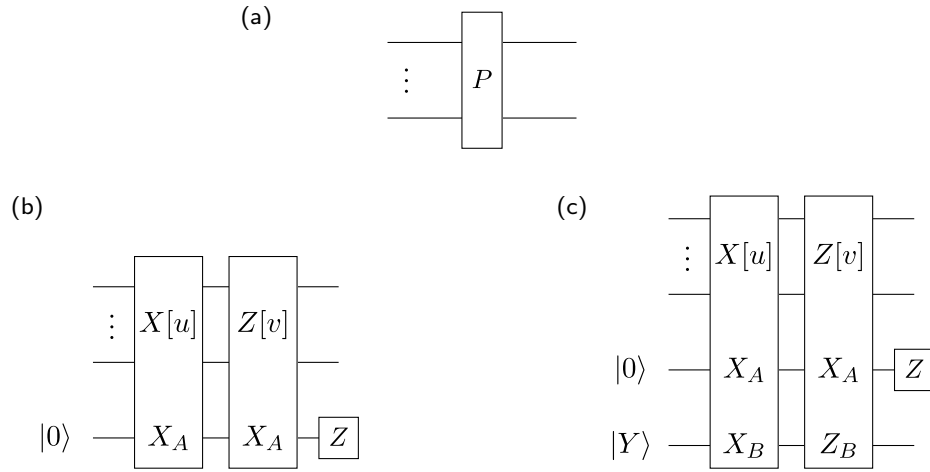


FIG. 10. Implementations of the Pauli product measurement $P = i^{u \cdot v} X[u]Z[v]$ from Ref. [49]. The measurement outcome of P is determined by the two split Pauli product measurements and the Z measurement. The $|Y\rangle$ state in (c) is left unchanged. (a) Pauli product measurement $P = iX[u]Z[v]$. (b) Equivalent measurement for $u \cdot v$ even. (c) Equivalent measurement for $u \cdot v$ odd.

While this is the same approach taken as in Ref. [47, Appendix H], we have two advantages. First, the overall space overhead of our scheme is lower as a consequence of using gauging logical measurements and universal adapters, rather than the CKBB scheme [28]. Second, this approach uses ancilla $|Y\rangle$ states, which are potentially costly to prepare in the context of [47]. The authors proposed to either use a qLDPC code that supports a transversal S gate or distill the $|Y\rangle$ states which incurs an additional overhead. The parallel measurement protocol we have proposed can prepare all required $|Y\rangle$ states in one measurement round. In other words, we bootstrap our measurement of arbitrary commuting sets off the ability of our scheme to measure Y operators without using $|Y\rangle$ states.

We now describe the procedure to measure arbitrary commuting products in more detail. We use the twist-free gadgets [65] from Ref. [49], which first appeared in Ref. [66]. In such gadgets, we split up a product into two steps using ancillae. This uses the fact that any Pauli product P to be measured can be described as $P = i^{u \cdot v} X[u]Z[v]$ for some binary vectors u and v defining the support of the X and Z components. If there are an even number of Y terms in P , so $u \cdot v$ is even, we initialize a single logical $|0\rangle$ state on qubit A and split P into two products: $X[u]X_A$ and $Z[v]X_A$. If there are an odd number of Y terms, so $u \cdot v$ is odd, we initialize one more logical $|Y\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle)$ state on qubit B and split P into $X[u]X_A X_B$ and $Z[v]X_A Z_B$. This $|Y\rangle$ state serves as a catalyst and is left unchanged by the measurements. These different cases are illustrated in Fig. 10.

Despite every product in Θ commuting, the decomposition above may produce products that anticommute. That anticommutation presents a problem for parallel

measurements: we would like to measure all the X -type products in the same timestep and then all the Z -type products. To ensure this does not happen, the operators in Θ must satisfy the following condition.

Definition 10 ([47, Definition 10]). The operator set Θ is said to be regular if and only if $u_i \cdot v_j = 0 \pmod 2$ for all Pauli products where $i \neq j$ in Θ .

Example 1. Let $\Theta = \{\bar{X}_1 \otimes \bar{Y}_2, \bar{Z}_1 \otimes \bar{X}_2\}$. Then

$$\begin{aligned} \bar{X}_1 \otimes \bar{Y}_2 &= \bar{X} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \bar{Z} \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\ \bar{Z}_1 \otimes \bar{X}_2 &= \bar{X} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \bar{Z} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \end{aligned}$$

ignoring phase factors. In this example Θ is not regular, as

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \neq 0, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \neq 0.$$

Although $[\bar{X}_1 \otimes \bar{Y}_2, \bar{Z}_1 \otimes \bar{X}_2] = 0$, $[\bar{X}_1 \otimes \bar{X}_2, \bar{Z}_1 \otimes \bar{Y}_2] \neq 0$, so the decompositions into X and Z -type products do not preserve commutation. As a consequence, the X -type products cannot be commuted through the Z -type products.

Given a general set Θ of commuting Pauli product measurements, Θ may not be regular. However, observing that Θ forms a generating set for a group of Pauli products $\langle \Theta \rangle$, it is shown in Ref. [47] that an alternative generating set $\Theta' \cup \Theta''$ of $\langle \Theta \rangle$ can always be chosen such that the subsets Θ' and Θ'' are each regular. Measuring Θ' and then Θ'' then performs the same logical operation on the code

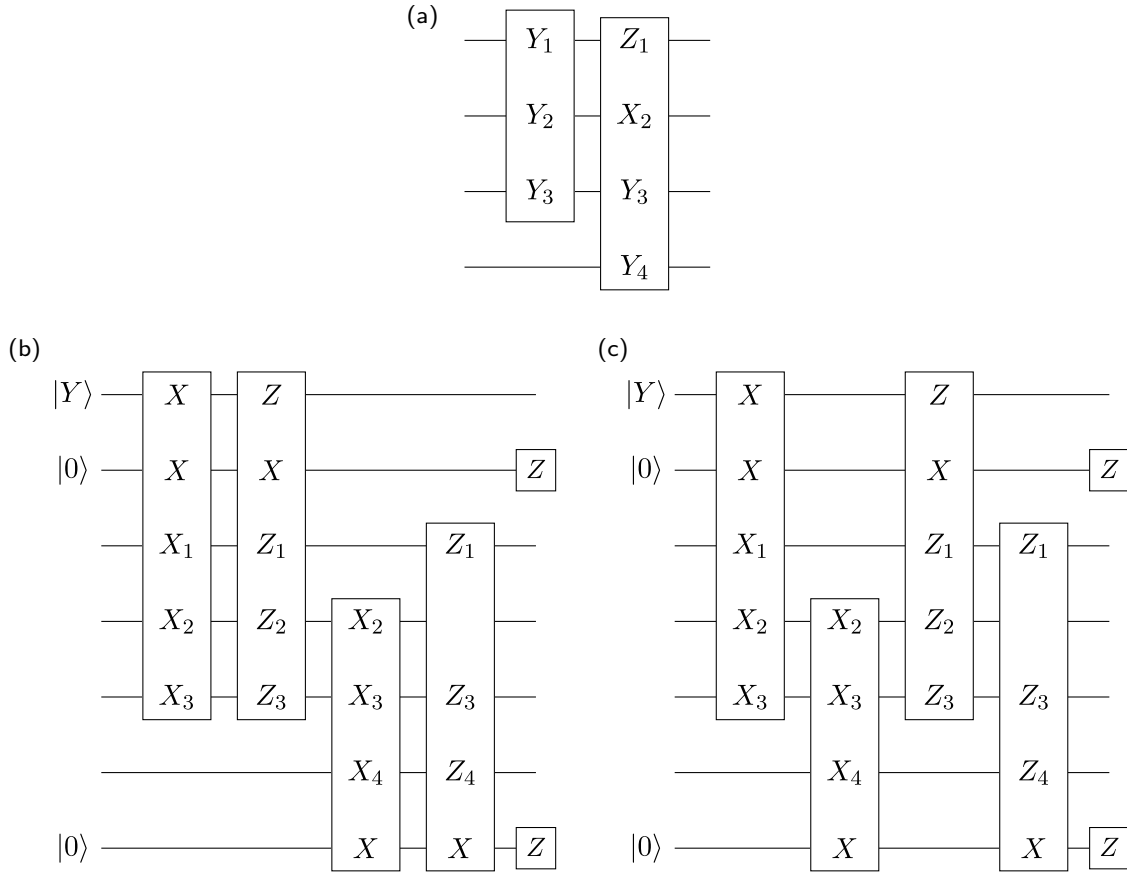


FIG. 11. Conversion of the regular set of Pauli product measurements $\Theta = \{Y_1 \otimes Y_2 \otimes Y_3, Z_1 \otimes X_2 \otimes Y_3 \otimes Y_4\}$ into twist-free surgeries, using the identities in Fig. 10. All X -type measurements are commuted to the start, which is possible because Θ is regular. The twist-free surgeries are implemented in two rounds of parallel measurements by surgery, not including the state preparation or final Z measurements. Note that in this case, the set Θ required only two logical rounds of measurement, as the set has only two Pauli product measurements, and so twist-free surgery does not reduce the time overhead. In general, commuting sets may contain up to k independent Pauli product measurements on k logical qubits. (a) Sequential measurement of $Y_1 \otimes Y_2 \otimes Y_3$ and $Z_1 \otimes X_2 \otimes Y_3 \otimes Y_4$. (b) Conversion into twist-free surgeries. (c) Commutation of all X -type measurements to the beginning.

as measuring Θ . We omit the details here, but refer the interested reader to [47, Lemma 9].

As this is a different set of Pauli product operators than Θ , the weights of the Pauli products in the set $\Theta' \cup \Theta''$ may differ substantially from those in Θ .

We then separate the measurement of our Pauli products into four steps:

- (1) Measure each $X[u_i]X_A$ or $X[u_i]X_AX_B$ for each $X[u_i]$ belonging to Θ' using a twist-free gadget.
- (2) Measure each $Z[v_i]X_A$ or $Z[v_i]X_AX_B$ for each $Z[v_i]$ belonging to Θ' using a twist-free gadget.
- (3) Measure each $X[u_i]X_A$ or $X[u_i]X_AX_B$ for each $X[u_i]$ belonging to Θ'' using a twist-free gadget.
- (4) Measure each $Z[v_i]X_A$ or $Z[v_i]X_AX_B$ for each $Z[v_i]$ belonging to Θ'' using a twist-free gadget.

We perform each step in parallel. Within each step, every Pauli product operator to be measured contains terms

which have the same or identity action on every logical qubit. As a consequence, we can use our previous branching and measurement scheme, introducing a leaf for each term to be measured, to perform these measurements in parallel. Each step requires $\mathcal{O}(t)$ $|0\rangle$ and $|Y\rangle$ ancilla logical states, and $\mathcal{O}(T\omega(\log T + \log^3 \omega))$ additional data qubits and checks to measure, where T is the total number of nonidentity terms of Pauli products in that step.

In Fig. 11 we show a circuit where Pauli product measurements are converted into twist-free surgeries, and the X -type measurements are commuted to the start, such that steps (1) and (2) may be performed. This separation into steps using twist-free surgery increases the time overhead of our scheme by a constant factor. Furthermore, the preparation of $\mathcal{O}(t)$ $|Y\rangle$ states, which can be reused throughout the computation as they are left unchanged by the twist-free surgery, can be performed in $\mathcal{O}(d)$ time by measuring $\tilde{Y}$ on each of those logical qubits at the

start of the computation. This removes a major obstacle in Ref. [47], where the preparation of these $|Y\rangle$ states incurs a significant overhead to distill or requires transversal S gates.

Finally, we remark that the ancilla logical $|0\rangle$ and $|Y\rangle$ states can be in the same codeblock on which we are performing logical measurements, or in separate codeblocks. The logical measurements involving these states are performed using the same branching and measurement scheme, described at the beginning of this section. The overhead of preparing these additional high-distance codeblocks is not a major concern, as we only need to prepare them once, which can be done using parallel measurement of individual $\bar{Y}$ logical operators, as specified in Lemma 9. Either way, we suffer a rate loss of at most a factor of 3, which is acceptable in the asymptotic regime.

IV. DISCUSSION

In this work we have presented a scheme for parallel quantum LDPC code surgery to fault-tolerantly measure logically disjoint Pauli products supported on t logical qubits using $\mathcal{O}(t\omega(\log t + \log^3 \omega))$ ancilla data qubits, and in time $\mathcal{O}(d)$, independent of t . This scheme requires no ancilla logical qubits. We then showed how to leverage this scheme to bootstrap a measurement protocol for arbitrary commuting sets of Pauli products, using $\mathcal{O}(t)$ ancilla logical qubits, and still not exceeding $\mathcal{O}(d)$ time to measure the commuting set.

Throughout, we assumed that branching and gauging measurement are separate steps, performed independently, to aid with the proofs of fault-distance in Appendix D. We expect that they can be performed simultaneously while maintaining the fault-distance, but have not proved this.

In practice, for small codes it has been found that decongestion is often not necessary, so the $\log^3 \omega$ factor vanishes from the overhead [40,41], and in general the upper bounds on overhead to preserve the distance tend to be very loose in quantum LDPC surgery [43]. Moreover, recent work implies that the $\log^3 \omega$ factor can be reduced to $\log \omega$ [67], although the precise details are yet to be worked out. The discussion in Ref. [67, Sec. 7.1.2] also gives evidence that our scheme may have optimal space overhead for a generic parallel logical measurement scheme, up to polylog factors, as our scheme can be seen as weight reduction: first of the checks required to measure the logical operators (gauging logical measurement), then of the qubits that those checks overlap on (brute-force branching).

It would be useful to avoid brute-force branching, to remove the $\log t$ factor and the formidable constants associated with branches, possibly by combining the gauging logical measurements framework with devised sticking [47], rather than brute-force branching. Unfortunately we could not find a construction for doing so which allowed us to provably use expander (hyper)graphs with gauged

cycles, and so retain a space overhead sub-quadratic in d . Reference [68] takes a similar approach, albeit without provable bounds on overhead. We remark that the use of the logical basis in standard form, along with toggling of connections using single-qubit Cliffords as in Lemma 4, allow for a different measurement scheme for logically disjoint Pauli products by devised sticking in $\mathcal{O}(\min(nd, t\omega d))$ space, an improvement on the $\mathcal{O}(t^2\omega d)$ of Ref. [47]. This would require moving to a subsystem code, however. We leave the details for future work.

If parallel LDPC code surgery is to be performed with branches, it is imperative to minimize the weights of logical representatives while maintaining suitable overlap properties. Another desirable improvement is to reduce the $\mathcal{O}(d)$ time overhead [69] of a logical round while retaining the parallelism of our scheme, without increasing space overhead substantially. Recent work on fast surgery accomplishes the reduction to $\mathcal{O}(1)$ syndrome measurement rounds, but either at the expense of space [70] or addressability on logical qubits [71]. Interestingly, the addressable scheme of Ref. [70] has the same asymptotic spacetime overhead as the parallel measurement scheme in this work, suggesting that there may be fundamental bounds on the asymptotic spacetime volume of distance-preserving addressable logical measurements on general quantum LDPC codes.

Recent work has focussed on performing surgery on quantum LDPC codes in a fixed-connectivity setting [54,55]. There are aspects of the scheme presented in this work that function in a fixed-connectivity setting. For example, branches for each single-qubit individual logical Pauli operator could be constructed in advance, along with their expander graphs for gauging logical measurement, then branches deleted when required. However, the space overhead of doing this will be exorbitant, and including all the connectivity for all possible adapters even more so. How to attain a reasonable degree of parallelism in a fixed-connectivity regime, either at blocklengths of practical relevance for large quantum algorithms or in the asymptotic regime, without increasing space cost substantially remains an open question, which we leave for future work.

The most important outstanding question for practical purposes is how much effect the combination of brute-force branching and measurement has on the circuit-level threshold of the code. Discerning this requires substantial circuit-level noise simulations and will depend heavily on the syndrome extraction circuits and decoder, so we also defer this to future work.

ACKNOWLEDGMENTS

We used TikZit [72] to generate the Tanner graph figures in this paper. This work was partially completed while A.C. was employed at Quantinuum. A.C. is grateful for

their support and helpful discussions with colleagues there. A.C. is also grateful to Ben Ide for pointing out the standard form of logical representatives for stabilizer codes in Ref. [50]. Z.H. is supported by the MIT Department of Mathematics and the NSF Graduate Research Fellowship Program under Grant No. 2141064. Z.H. would like to thank Tomas Jochym-O'Connor for pointing out Refs. [62,63]. D.W. is currently on leave from The University of Sydney and is employed by PsiQuantum. Part of this work was initiated when Z.H., A.C., and D.W. were visiting the Fault-Tolerant Quantum Technologies Workshop in Benasque, 2024. We are grateful to the organizers of the workshop.

DATA AVAILABILITY

No data were created or analyzed in this study.

APPENDIX A: CONNECTIONS TO HOMOLOGICAL ALGEBRA

1. Hypergraph products

Hypergraph products [57] are essentially tensor products of chain complexes, where one of the complexes has been dualized. That is, denoting the hypergraph product $[C, D]_\bullet$ [73] then

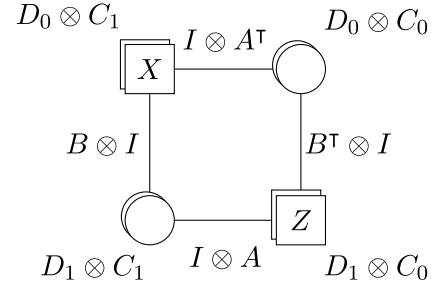
$$[C, D]_\bullet := (D \otimes C^*)_\bullet,$$

where $- \otimes -$ is the tensor product of chain complexes [74,75].

Given $C_\bullet$ and $D_\bullet$ as scalable Tanner graphs,



which are classical codes and so have no Z checks (in our convention), then the hypergraph product $[C, D]_\bullet$ is

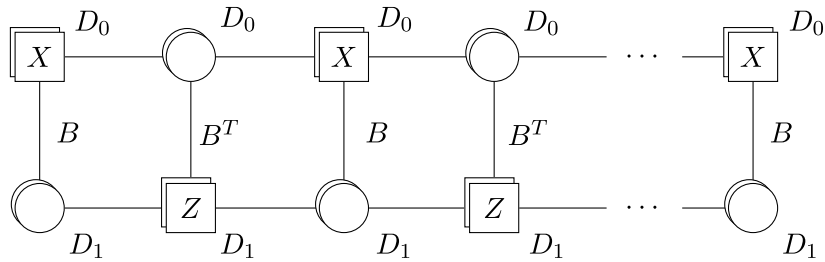


in scalable Tanner graph notation.

In particular, if we take the repetition code $R_\bullet = R_1 \rightarrow R_0$, with parity-check matrix

$$\begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

then the scalable Tanner graph of $[R, D]_\bullet$ is



where all horizontal edges carry the label I , i.e., they are identity maps. We can see that if $D_\bullet$ is the classical code corresponding to a $\bar{Z}$ logical in a larger code Q , and its incident X checks, then gluing $[R, D]_\bullet$ into Q as a branching sticker makes another equivalent logical on each odd layer of this scalable Tanner graph; in particular, there is an equivalent logical on the very last layer. In brute-force branching, we initialize trees of such hypergraph products where each branch uses $R_\bullet$ with the check matrix $\begin{pmatrix} 1 & 1 \end{pmatrix}$, but varying $D_\bullet$ to have a classical code decreasing in size, to give equivalent logicals without overlap, as in Fig. 6.

2. Branching as a series of mapping cylinders

When the initial code Q is CSS, brute-force branching of all Z or X -type operators can be rephrased in the language of chain complexes. This appendix assumes the reader is familiar with elementary homological algebra. By standard theory [6], Q can be viewed as a chain complex $C_\bullet$ over $\mathbb{F}_2$,

$$C_\bullet = C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0,$$

where $\partial_2 = H_Z^T$ and $\partial_1 = H_X$, and each vector space is assigned a basis, such that

$$C_2 = \mathbb{F}_2^{m_z}, \quad C_1 = \mathbb{F}_2^n, \quad C_0 = \mathbb{F}_2^{m_x},$$

where n , m_x , and m_z are the number of data qubits, X -type and Z -type stabilizer generators, respectively. Each data qubit in Q can be identified with a basis element in C_1 , and $X(Z)$ -checks can be identified with basis elements in $C_0(C_2)$, respectively. The number of logical qubits in Q is determined by the homology of $C_\bullet$ at degree 1, i.e.,

$$k_Q = \dim H_1(C).$$

Furthermore, each $\bar{Z}$ logical operator can be identified with a vector $v \in C_1$, such that $v \in \ker \partial_1$; for each nontrivial $\bar{Z}$ logical operator, $v \in \ker \partial_1 \setminus \text{im} \partial_2$. Each equivalence class of operators (up to Z stabilizers) with representative $\bar{Z}$ is associated with $[v]$, the equivalence class in $H_1(C) = \ker \partial_1 / \text{im} \partial_2$.

Let $\{\bar{Z}_i\}_{i \in I}$ be a set of $\bar{Z}$ logical operators to branch. $\{\bar{Z}_i\}_{i \in I}$ can be viewed as a set of $\mathbb{F}_2$ -vectors $v_i \in C_1$.

Let $A_1 \subset C_1$ be a subspace of C_1 , defined by $A_1 = \bigcup_i \text{supp}(v_i)$. That is, a basis element of C_1 is included in A_1 if and only if it is in the support of a $\bar{Z}$ operator to be branched.

Then define a matrix $\partial_1^A := \partial_1|_{A_1}$, restricting H_X to its support on A_1 . The codomain of this matrix is $A_0 = \bigcup_{u \in \text{im} \partial_1^A} \text{supp}(u)$. In other words, A_0 has as basis elements all X checks which are incident to any qubits in the support of any $\bar{Z}$ operators to be branched.

The complex $A_\bullet = A_1 \rightarrow A_0$ is a subcomplex of $C_\bullet$, i.e., there is an injective chain map $f_\bullet : A_\bullet \hookrightarrow C_\bullet$, which maps basis elements to basis elements. As a commutative diagram,

$$\begin{array}{ccccc} 0 & \longrightarrow & A_1 & \xrightarrow{\partial_1^A} & A_0 \\ \downarrow f_2 & & \downarrow f_1 & & \downarrow f_0 \\ C_2 & \xrightarrow{\partial_2} & C_1 & \xrightarrow{\partial_1} & C_0 \end{array}.$$

Note that the zero object and f_2 map can be omitted, as the map must always be zero. As f_1 and f_0 map basis elements to basis elements, each of them has the form

$$\begin{pmatrix} I \\ 0 \end{pmatrix}$$

up to permutation.

The branch can then be constructed as $\text{cyl}(f)$, where $\text{cyl}(\cdot)$ is the mapping cylinder [75, Sec. 1.5]. Diagrammatically,

$$\text{cyl}(f) = \begin{array}{ccccc} & & A_1 & \xrightarrow{\partial_1^A} & A_0 \\ & \nearrow I & & \nearrow I & \\ A_1 & \xrightarrow{\partial_1^A} & A_0 & & \\ & \searrow f_1 & & \searrow f_0 & \\ C_2 & \xrightarrow{\partial_2} & C_1 & \xrightarrow{\partial_1} & C_0 \end{array}$$

where direct sums are taken vertically on the vector spaces.

To orient this construction in terms of quantum codes and Sec. II B, observe that for any $\bar{Z}_i$ with $i \in I$, multiplying by the appropriate element in $\text{im} \partial_1^{\text{cyl}(A)}$ will shift the support v_i from its location in the original code to the new data qubits A_1 on the top row. That is, there is now an equivalent representative of $\bar{Z}_i$ with support wholly outside of the original code, and so the mapping cylinder creates a branch.

By inspection there is an injective chain map $\alpha_\bullet : C_\bullet \hookrightarrow \text{cyl}(f)$, mapping $C_i \rightarrow C_i$.

Lemma A1 ([75], Lemma 1.5.6). The inclusion $\alpha_\bullet : C_\bullet \hookrightarrow \text{cyl}(f)$ is a quasi-isomorphism.

That is, the restriction and corestriction of $\alpha_\bullet$ to homology spaces is an isomorphism, i.e., $H_i(\alpha)$ is an isomorphism for all degrees i .

As a consequence, the addition of a branch introduces no new logical operators. Note that the above Lemma is stronger than $H_1(C) \cong H_1(\text{cyl}(f))$ as vector spaces; it also dictates that this isomorphism is given by $H_1(\alpha)$, which means that every $\bar{Z}$ logical operator in $C_\bullet$ is also present in $\text{cyl}(f)$ and the equivalence classes are preserved, so it is not the case that some logical qubits are removed and replaced with different ones. See, e.g., Ref. [42] or Ref. [76, Sec. 2.2.3] for discussion on interpreting logical maps, i.e., maps between homology spaces, from physical maps, i.e., maps between chain complexes.

Next, one can make a branch for a different set $\{\bar{Z}_j\}_{j \in J}$ of $\bar{Z}$ logical operators, in the same manner [77]. This second

branch is still well-defined by checking the diagram,

$$\text{cyl}(f + g) = \begin{array}{ccccc} & & A_1 & \xrightarrow{\partial_1^A} & A_0 \\ & \nearrow I & & \nearrow I & \\ A_1 & \xrightarrow{\partial_1^A} & A_0 & & \\ & \searrow f_1 & & \searrow f_0 & \\ C_2 & \xrightarrow{\partial_2} & C_1 & \xrightarrow{\partial_1} & C_0 \\ & \nearrow g_1 & & \nearrow g_0 & \\ B_1 & \xrightarrow{\partial_1^B} & B_0 & & \\ & \searrow I & & \searrow I & \\ & & B_1 & \xrightarrow{\partial_1^B} & B_0 \end{array}$$

By iteration, one can build up a series of mapping cylinders on $A_\bullet$ and $B_\bullet$, branching all $\bar{Z}$ logical operators to have leaves. By applying Lemma A1 sequentially, one immediately acquires a series of quasi-isomorphisms,

$$C_\bullet \xrightarrow{\sim} \text{cyl}(f) \xrightarrow{\sim} \text{cyl}(f + g) \xrightarrow{\sim} \dots$$

and so brute-force branching introduces no new logical operators. This yields an alternative, more algebraic, proof of Lemma 1.

Remark A1. In Ref. [51, Appendix A] mapping cylinders were suggested as a potential formalism for treating joint measurement. Here we use them for branching, a different but related application.

Remark A2. We claim that the same proof via mapping cylinders also applies to non-CSS stabilizer codes, but this requires including the symplectic form in the chain complexes explicitly [78], which complicates the construction. In our case, all the branching on non-CSS codes toggles check type to make the code locally CSS, see Lemma 4, which obviates the need for such a formalism for branching, as Lemma 1 relies only on the code locally around the representatives being branched.

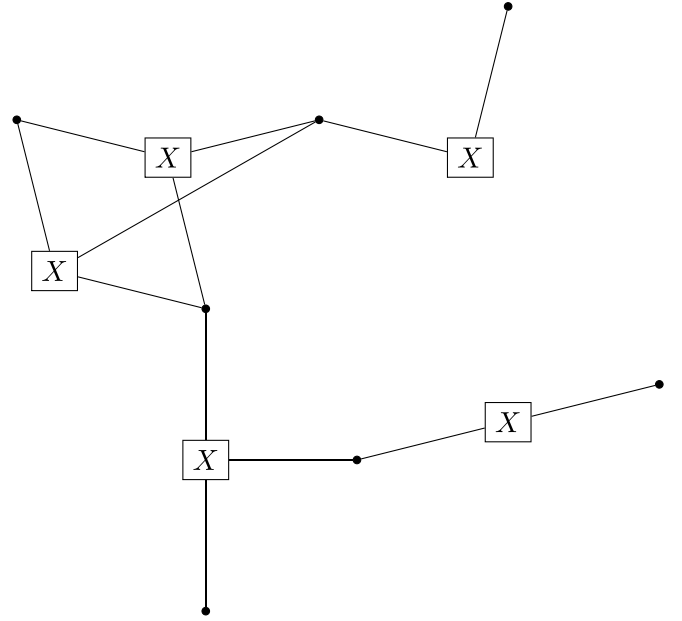
APPENDIX B: TOY EXAMPLES OF BRUTE-FORCE BRANCHING

In this appendix we give explicit examples of brute-force branching and measurement.

1. Brute-force branching $\bar{Z}$ logicals

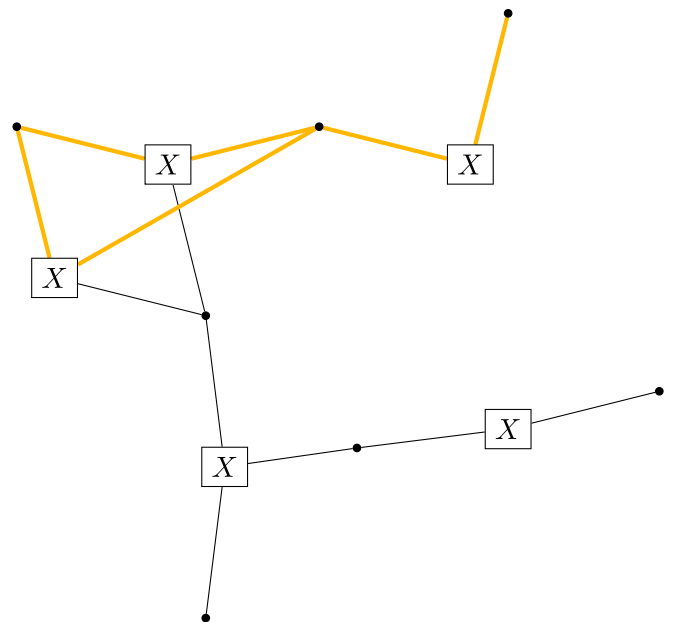
We start with a toy set of $\bar{Z}$ logicals. Consider a CSS code Q , with Tanner graph $T(Q)$, such that there is a

Tanner subgraph $\mathcal{G}(V_{\text{data}} \cup V_{\text{check}}, E)$,



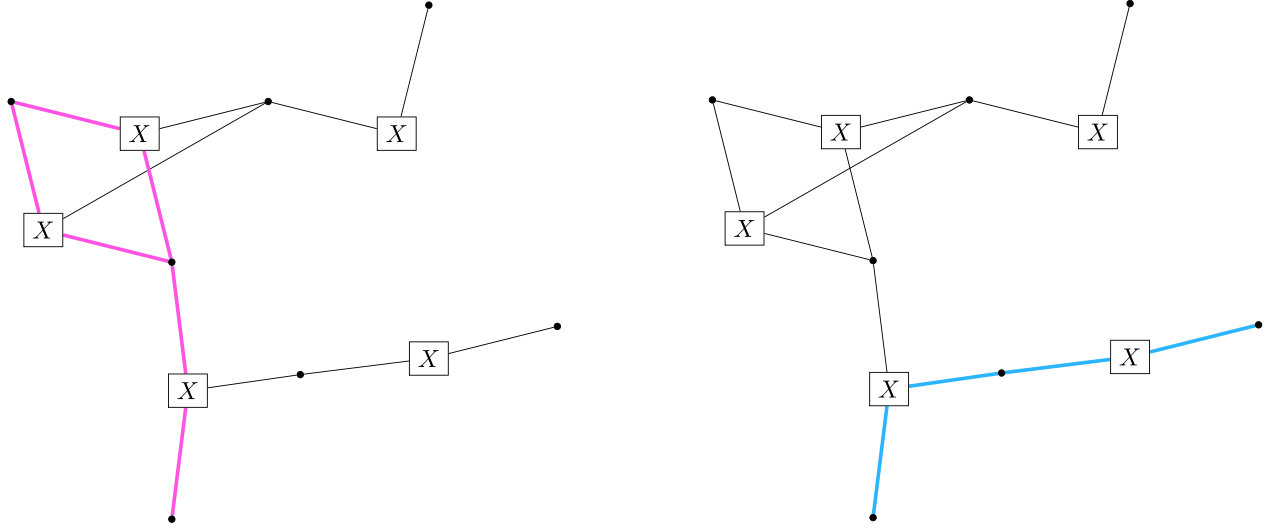
For simplicity of presentation, we do not consider the rest of the Tanner graph $T(Q)$ (and therefore do not specify the actual initial code Q). Crucially, the X -checks in $\mathcal{G}$ are the *only* X -checks incident to the data qubits in V_{data} ; otherwise, one can check that the deformed code would not generally have commuting stabilizers. The data qubits in V_{data} may be incident to Z -checks elsewhere in $T(Q)$, and the X -checks in V_{check} may be incident to other data qubits in $T(Q)$.

Observe that there are three independent $\bar{Z}$ logicals in $\mathcal{G}$: v_1 , v_2 , and v_3 . We highlight v_1 in orange,

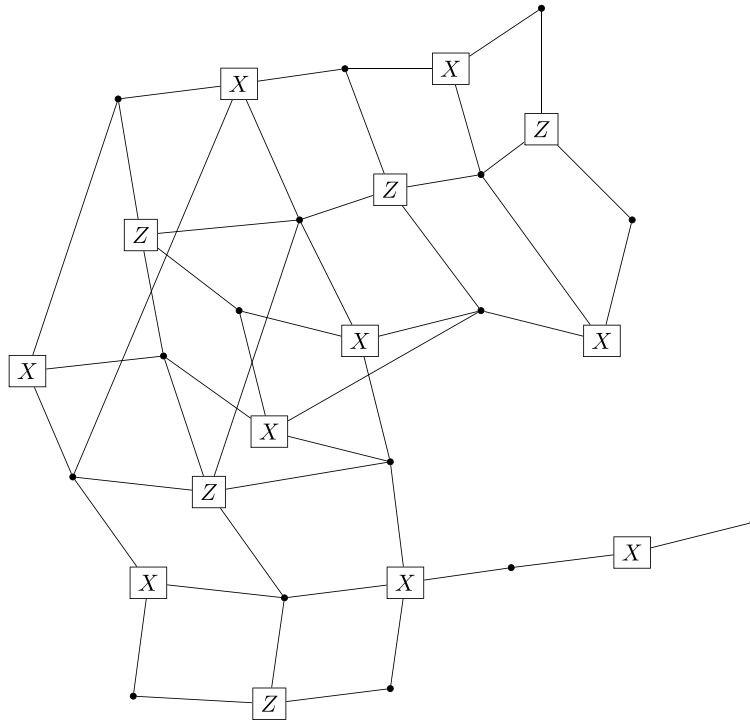


v_2 in pink,

and v_3 in blue.

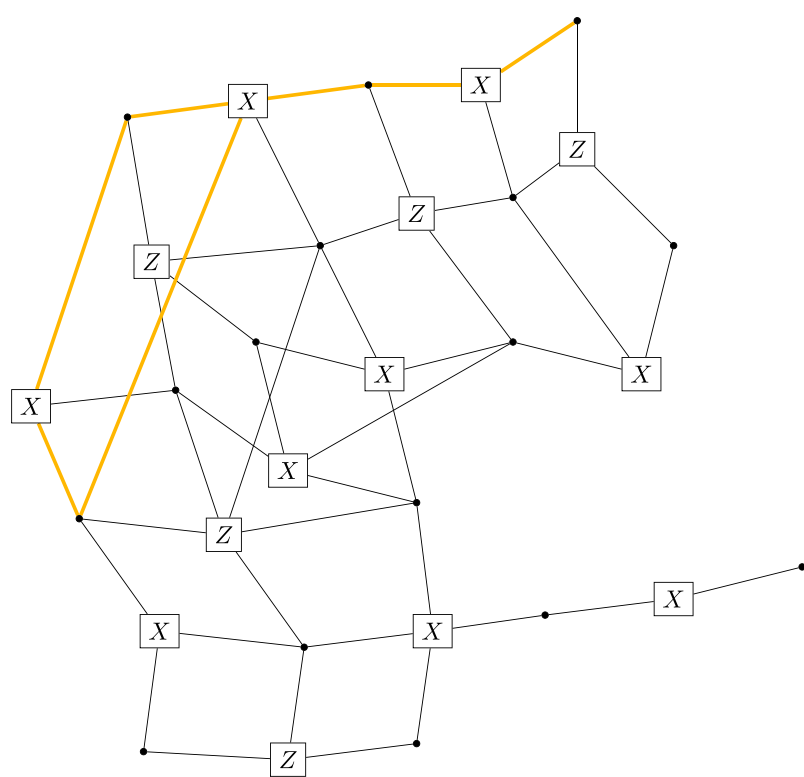


Each of the $\bar{Z}$ logicals in $\mathcal{G}$ have weight at least 3, including those that are linear combinations. Following the description in Sec. II B, let $\mathcal{I} = \{v_1, v_2, v_3\}$, and $t = 3$. We choose $\lceil t/2 \rceil = 2$ logicals from $\mathcal{I}$, say v_1 and v_2 . We say that $G_1(S_1 \cup \chi_1, E_1)$ is the subgraph of $\mathcal{G}$ where $S_1 = \text{supp}(v_1) \cup \text{supp}(v_2)$ and χ_1 is the set of X -checks incident to any vertex in S_1 . Then the branch for the subgraph G_1 has the following form:

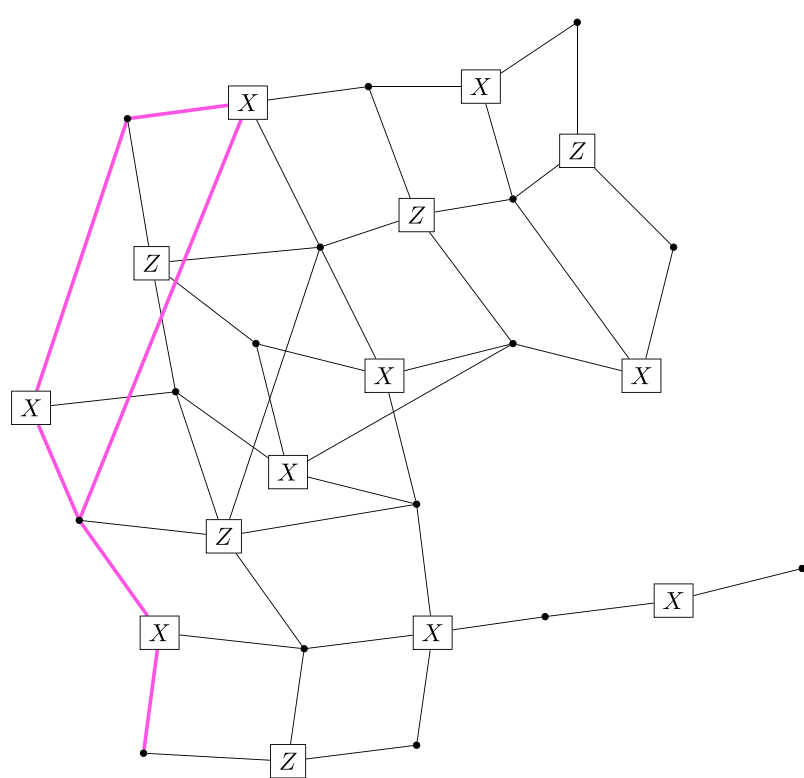


Call this Tanner graph $B_1(\mathcal{G})$. Each data qubit in the support of S_1 has a new Z -check in $S_1^\top$ connected to it; each X -check in χ_1 has a new data qubit in $\chi_1^\top$ connected to it, and these have connections between them which mirror G_1 . Then, each new Z -check in $S_1^\top$ has another new data qubit connected to it, and each new data qubit in $\chi_1^\top$ has another new X -check connected to it, in H_1 , the copy of G_1 , which is placed on the top left. Note that there is an injection of graphs $\mathcal{G} \subset B_1(\mathcal{G})$, i.e., we have built on top of the previous subgraph $\mathcal{G}$. In $B_1(\mathcal{G})$ each of the logicals v_1, v_2 can be cleaned to the end of the branch.

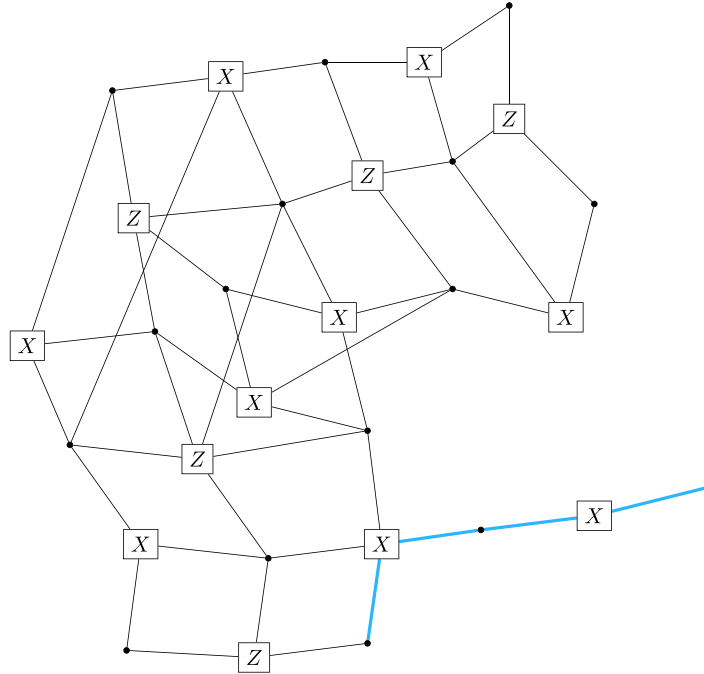
v'_1 , which is v_1 after cleaning, is highlighted in orange below,



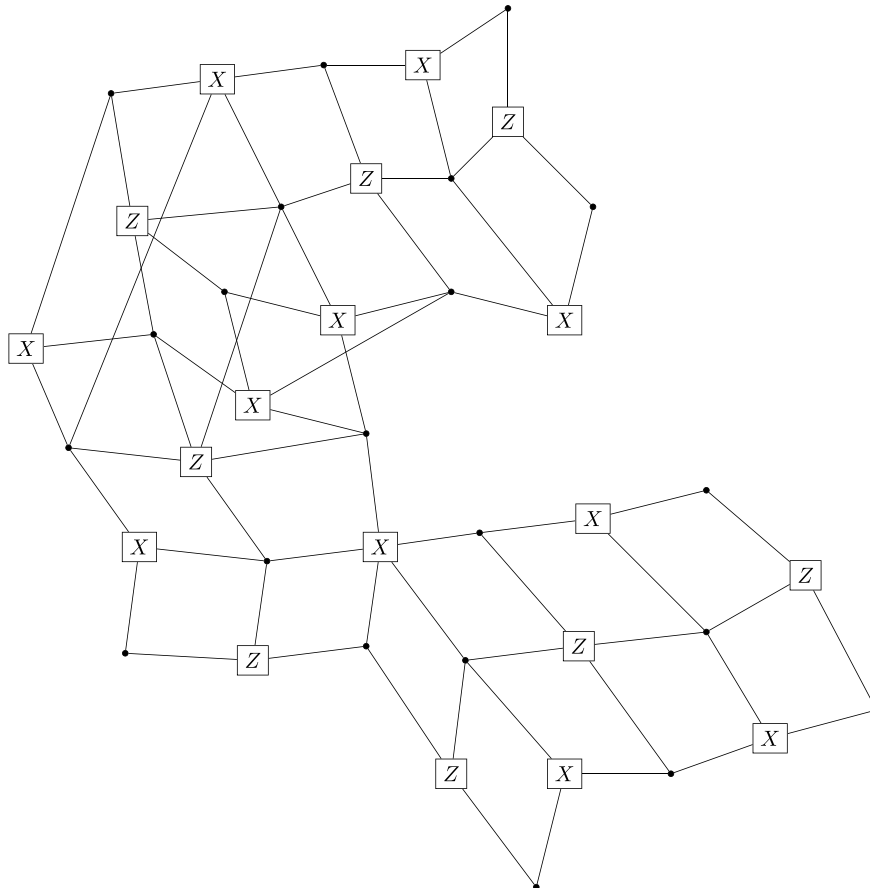
and v'_2 is highlighted in pink,



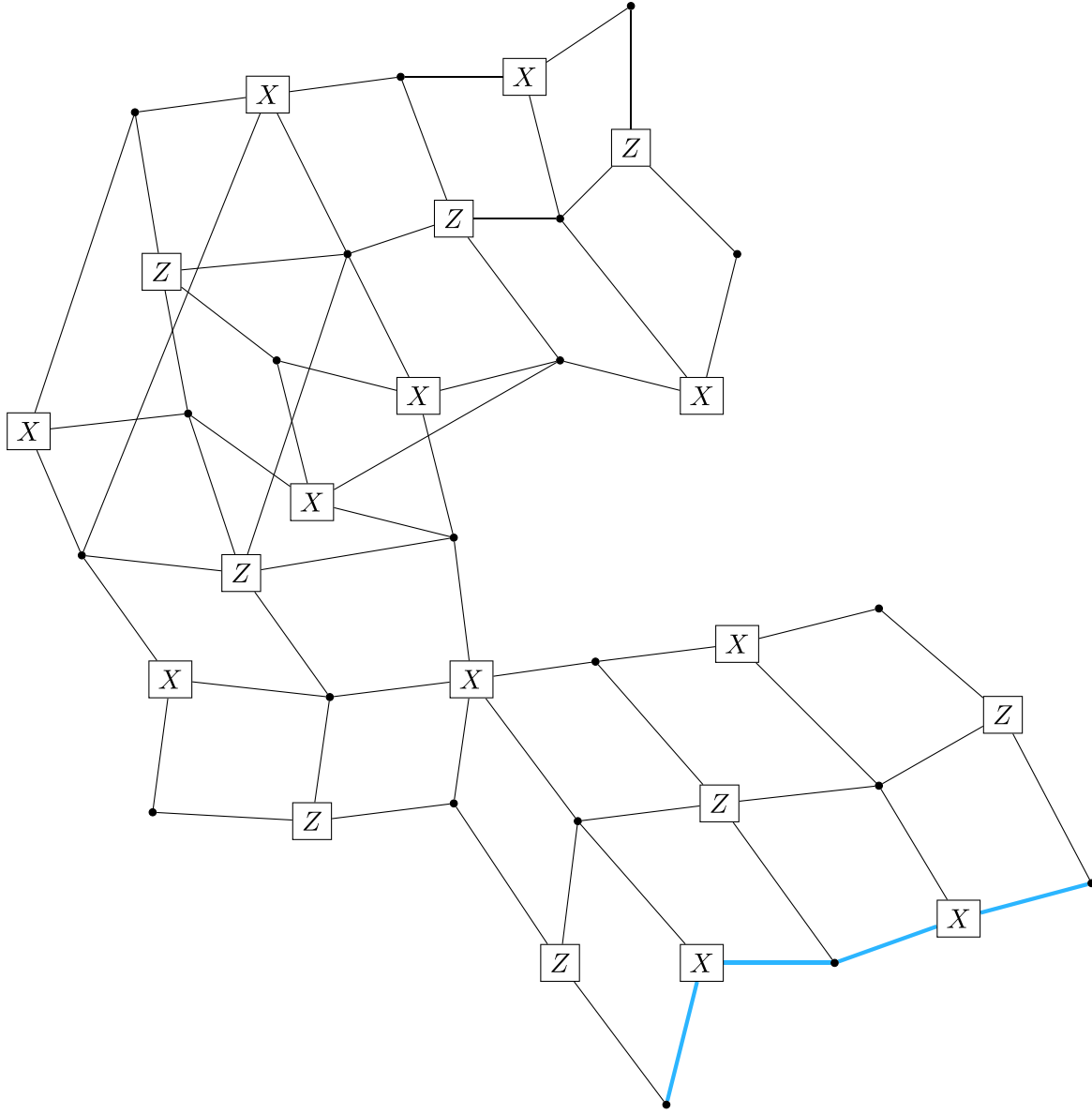
Now, v'_1 and v'_2 are not disjoint from one another, but are disjoint from v_3 :



As v_3 , the other $\lfloor t/2 \rfloor = 1$ operator, is now disjoint from v'_1 and v'_2 , we could leave v_3 unchanged, but it is illustrative to branch v_3 as well, to indicate the structure for $t > 3$. Repeating the branching procedure for v_3 yields the following Tanner graph:



where v'_3 is highlighted in blue below:



One could then continue, branching out v'_1 and v'_2 , which we omit. This explicit Tanner graph is an instantiation of the more abstract illustration in Fig. 5 before the last round of branching, which would yield disjoint representatives v''_1, v''_2 .

Immediately from this toy example, one can see that for low-distance or low-rate codes, branching may incur a substantial practical cost. However, for codes which are high-distance and high-rate, branching followed by gauging measurement and adapters is asymptotically efficient.

2. Parallel measurement of mixed logicals

We now give a toy example of brute-force branching and measuring mixed logicals. For this example, we use the $[[8, 3, 3]]$ non-CSS stabilizer code from Ref. [50, Table 3.3].

The $[[8, 3, 3]]$ code has the stabilizer matrix,

$$\left[\begin{array}{cccccccc|cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{array} \right].$$

As there are checks with weight 8, the same as the length of the code, there is not a great deal of point in performing LDPC surgery on the $[[8, 3, 3]]$ code, as there is no sparsity to be preserved. Nevertheless, it is a reasonable toy example for the purpose of demonstration.

In standard form [50, Sec. 4.1], the logical operators have the structure

$$\left[\begin{array}{c|cccccccc|cccccccc} \bar{X}_1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ \bar{X}_2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ \bar{X}_3 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \bar{Z}_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ \bar{Z}_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ \bar{Z}_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{array} \right].$$

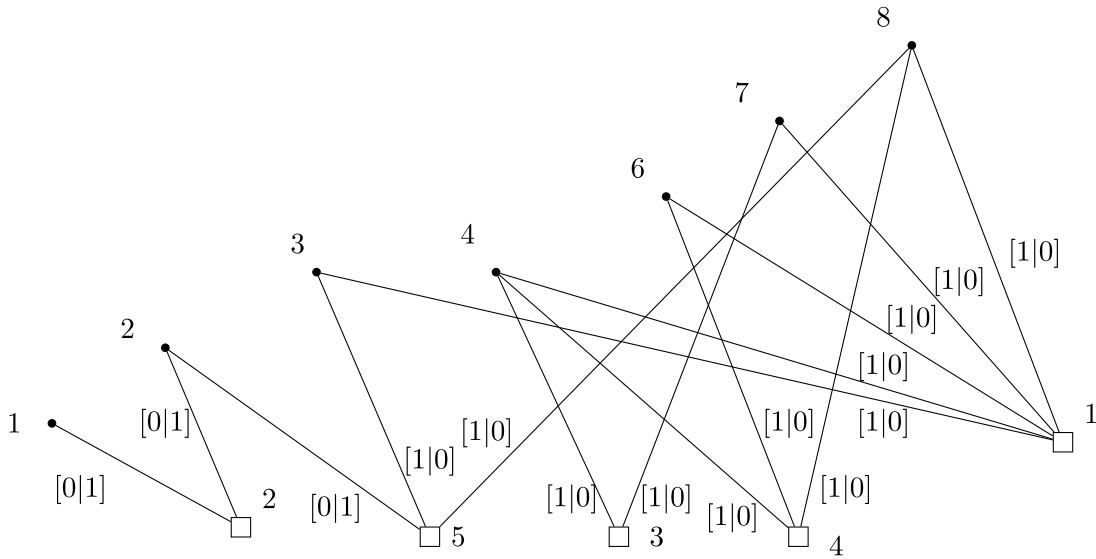
We choose an arbitrary selection of logical operators on disjoint logical qubits to measure in parallel: $\bar{X}_1, \bar{Z}_2, \bar{Y}_3$. As they act on disjoint logical qubits, and the logical basis is in standard form, by Lemma 4 we can branch simultaneously.

For illustrative purposes, we shall partition the terms into $\{\bar{X}_1, \bar{Z}_2\}$ and $\{\bar{Y}_3\}$, and branch $\{\bar{X}_1, \bar{Z}_2\}$ together, then $\{\bar{Y}_3\}$. Then we shall split $\{\bar{X}_1, \bar{Z}_2\}$ onto leaves and finally apply the gauging logical measurement scheme to measure $\bar{X}_1, \bar{Z}_2, \bar{Y}_3$.

The union of $\bar{X}_1$ and $\bar{Z}_2$ acts as

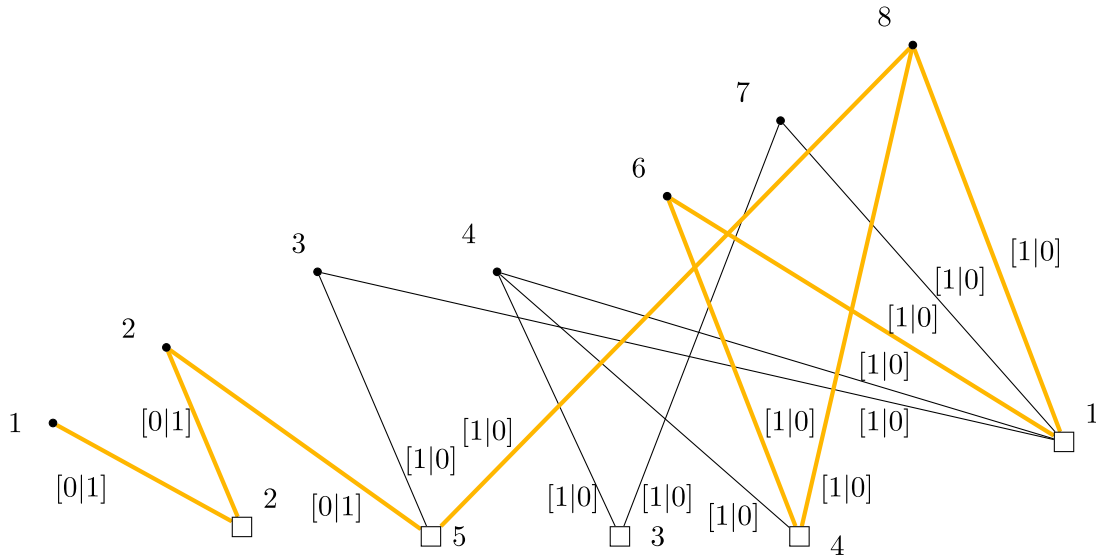
$$\left[\bar{X}_1 \cup \bar{Z}_2 \mid 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \mid 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \right].$$

Every check in the code has a component which anticommutes with the action of $\bar{X}_1 \cup \bar{Z}_2$, so the first branch will introduce five new data qubits, as well as the seven new checks. The local structure of the Tanner graph $\mathcal{T}(Q)$ around $\bar{X}_1 \cup \bar{Z}_2$ is

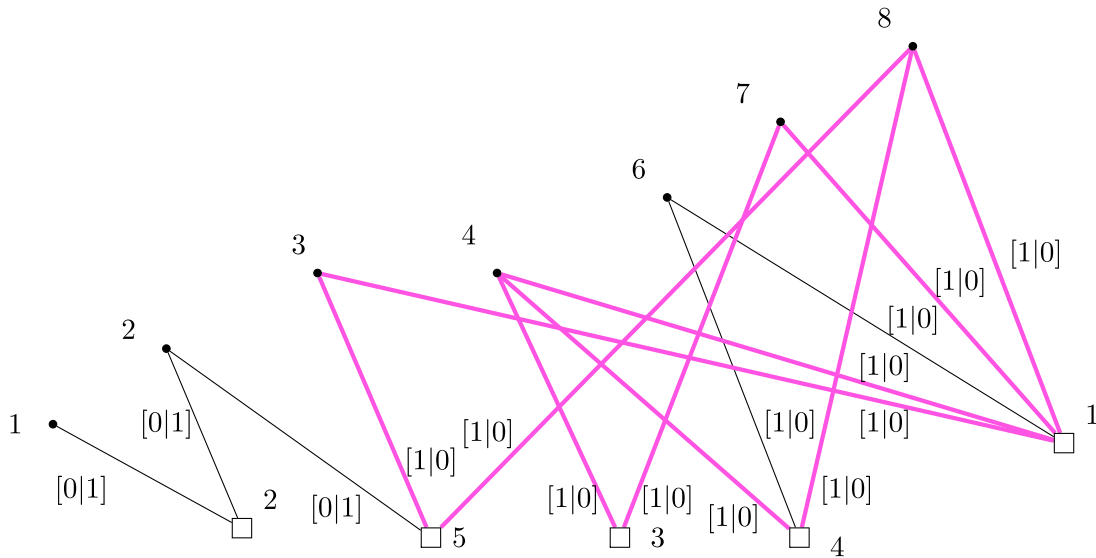


where vertices are labeled according to columns and rows in the check matrix, and edges are labeled according to Definition 5.

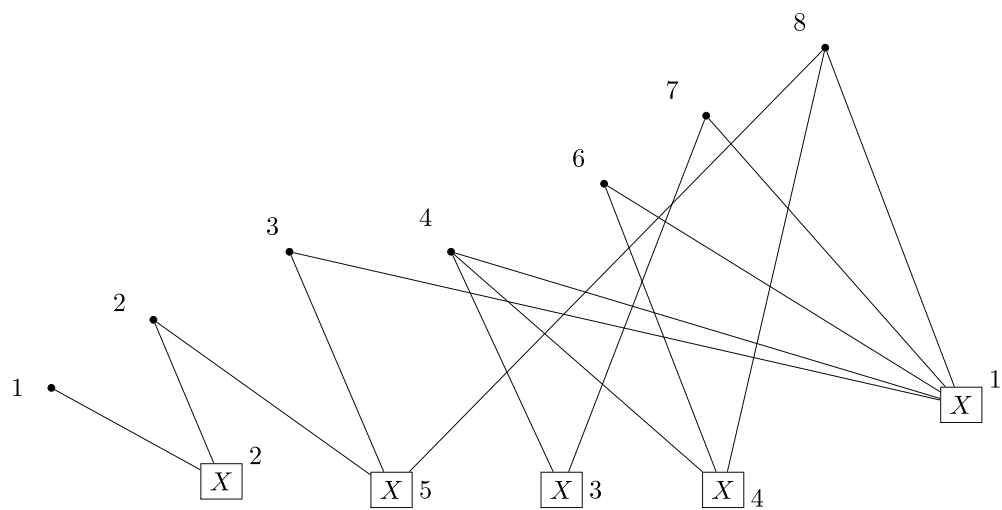
$\bar{X}_1$ is highlighted in orange,



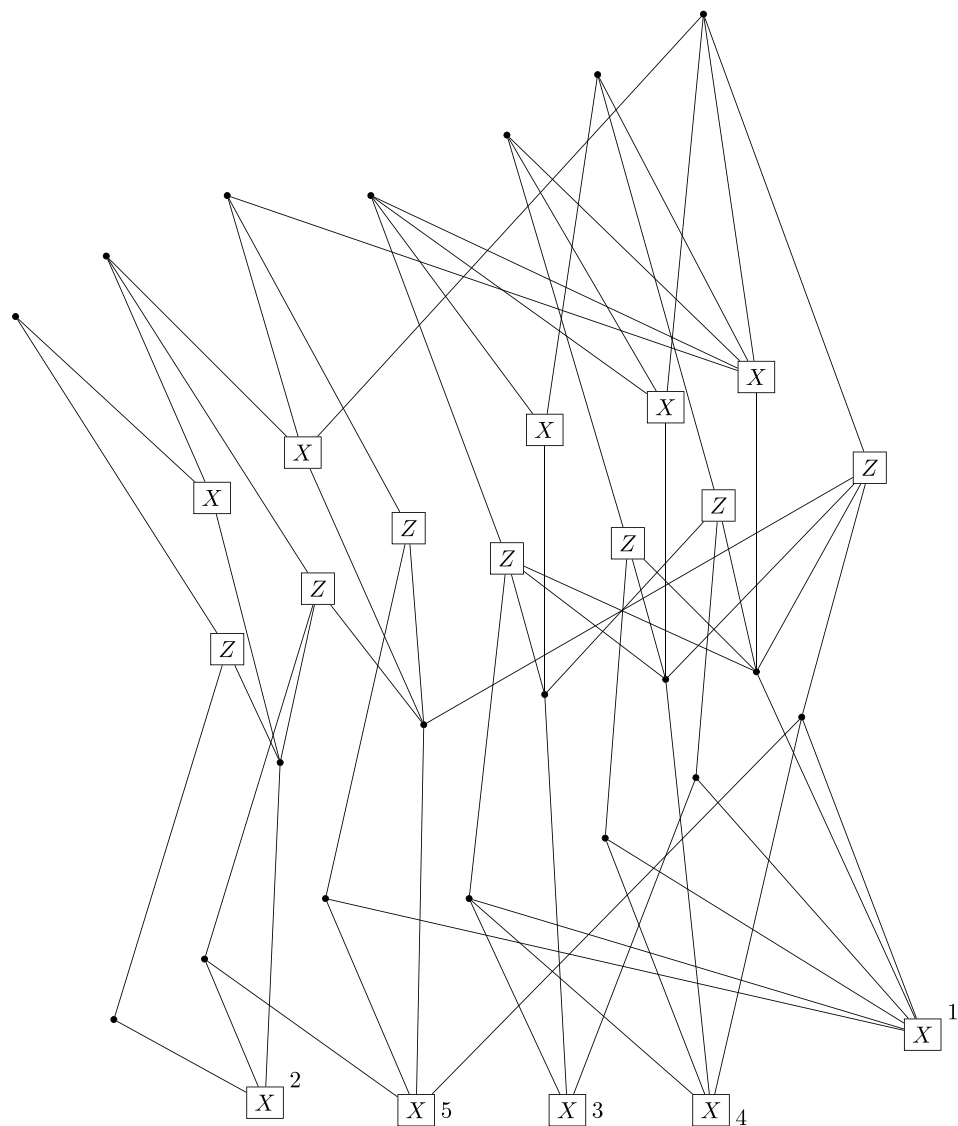
and $\bar{Z}_2$ in pink,



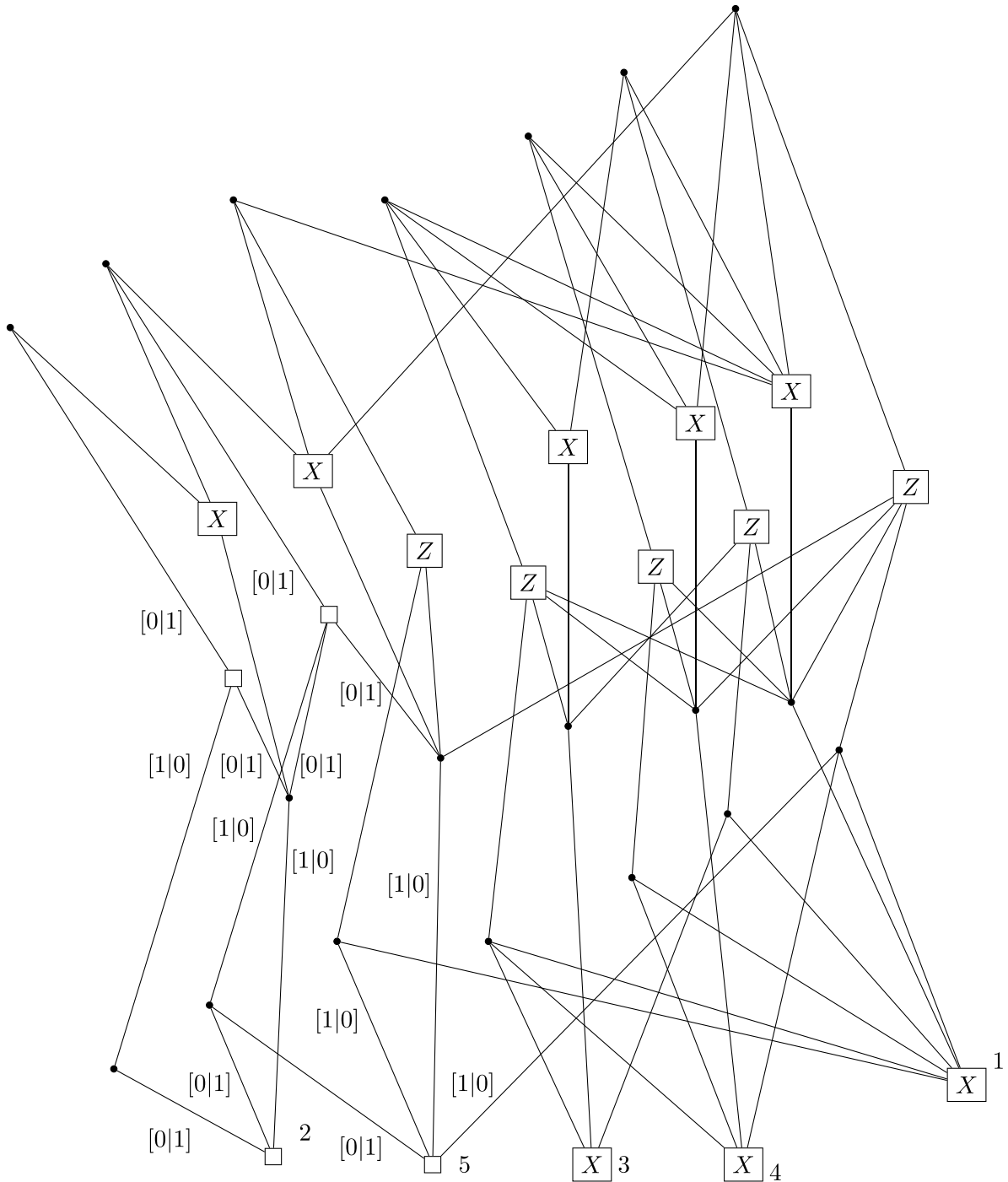
As the X -component of $\bar{X}_1$ is disjoint from the Z -component of $\bar{Z}_2$, and vice versa, we can toggle the X -component of the union, on qubits 1 and 2, to act as Z instead, using single-qubit Cliffords (in this case Hadamards), which yields the local Tanner graph,



In other words, despite the whole code not being CSS, we can consider the local code around the logicals to be branched to be CSS and define a subcomplex as in [Appendix A 2](#).
Using Tanner graphs, the branch has the following form:



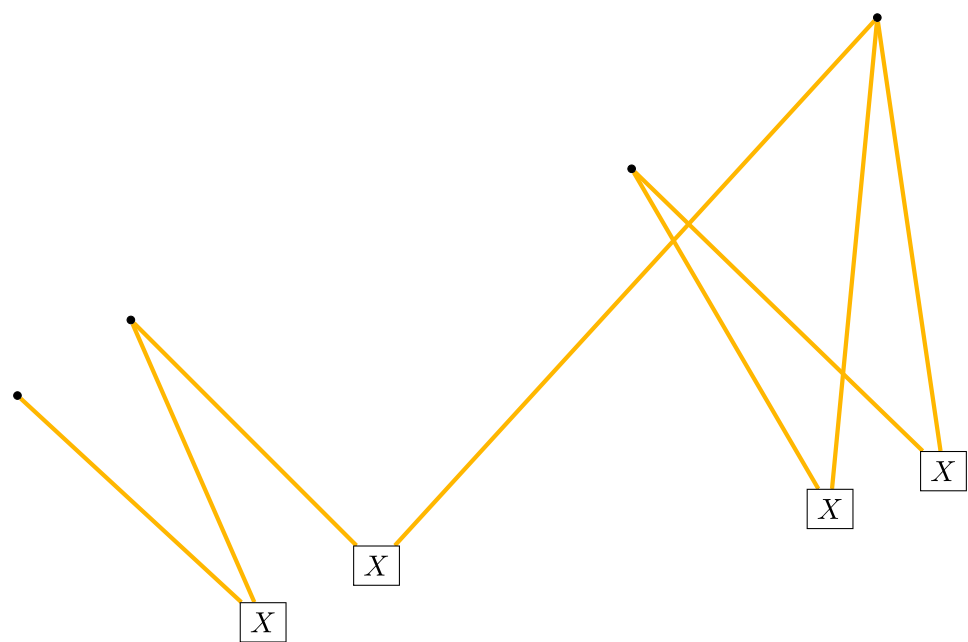
Afterwards, to make the local code agree with the rest of Q again, we toggle the X -component of the union in Q back.



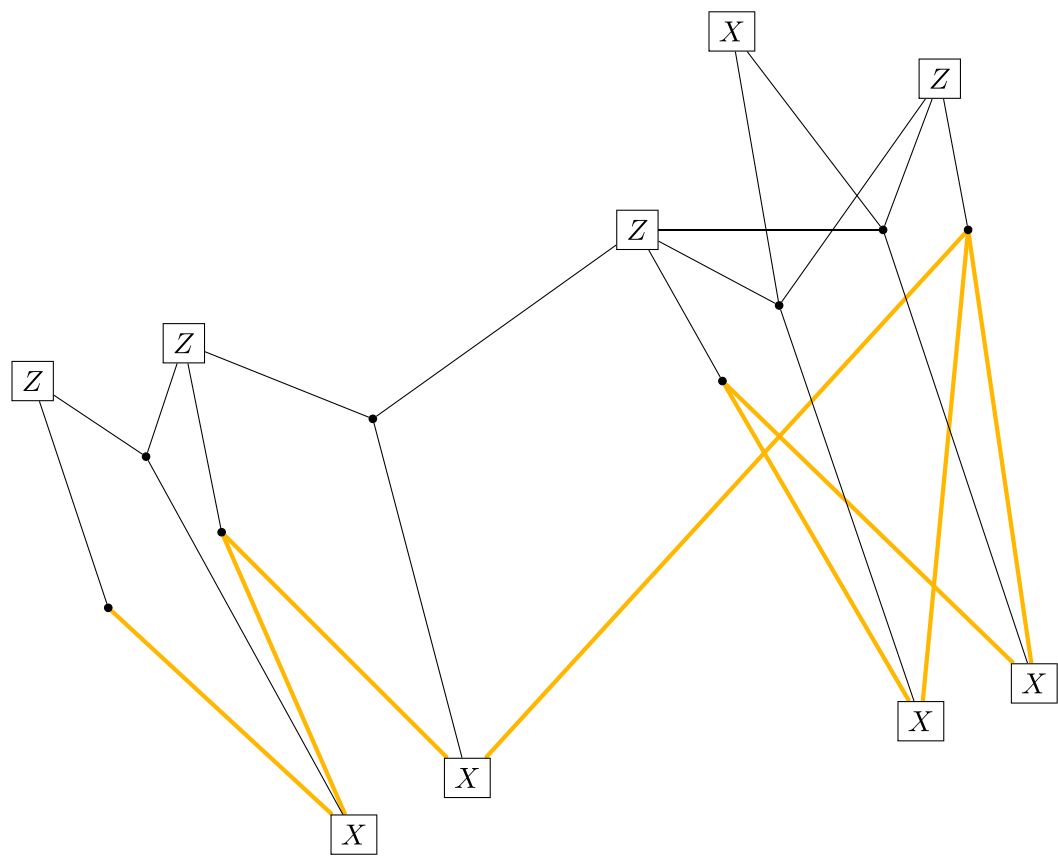
Observe that after cleaning to the end of the branch, both $\bar{X}_1$ and $\bar{Z}_2$ act as products of Z Paulis.

Lastly, we can branch $\bar{X}_1$ and $\bar{Z}_2$ to be on distinct leaves, or measure them in-place. The same goes for $\bar{Y}_3$.

Focussing first on $\tilde{X}_1$,

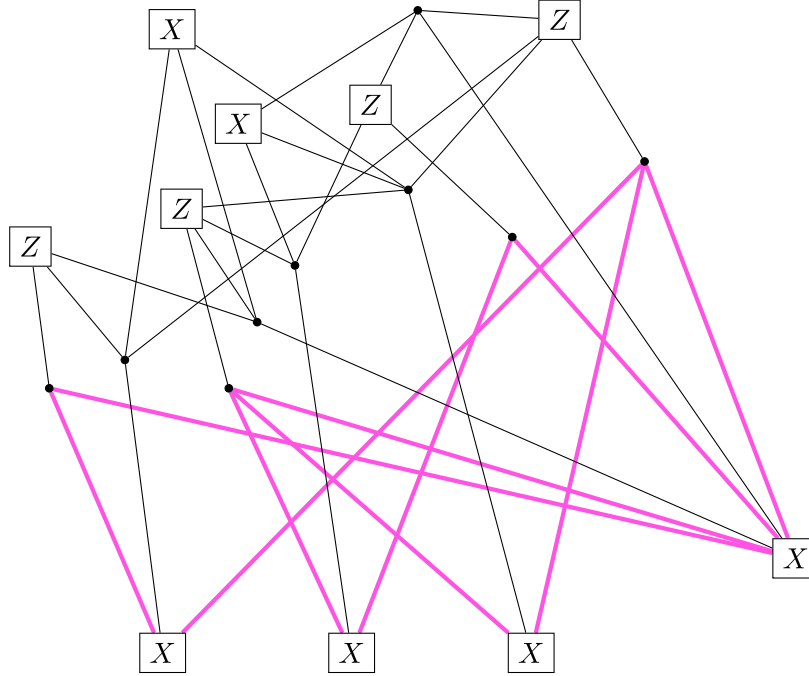


we follow Ref. [40] and construct a measurement graph,



where new Z -checks are vertices in the measurement graph, new data qubits are edges in the measurement graph, and the new X -check gauge-fixes the cycle in the graph, cf. Definition 6.

Repeating the process for $\bar{Z}_2$, we have



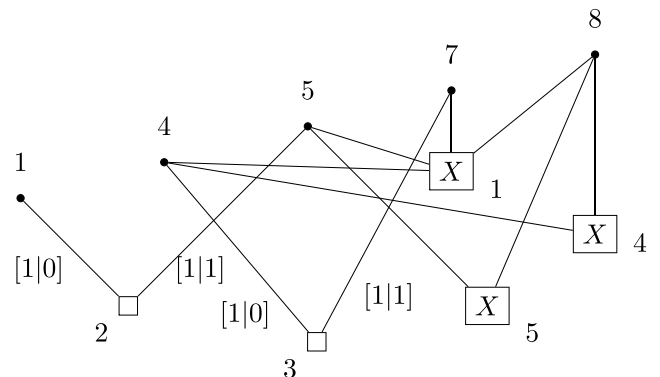
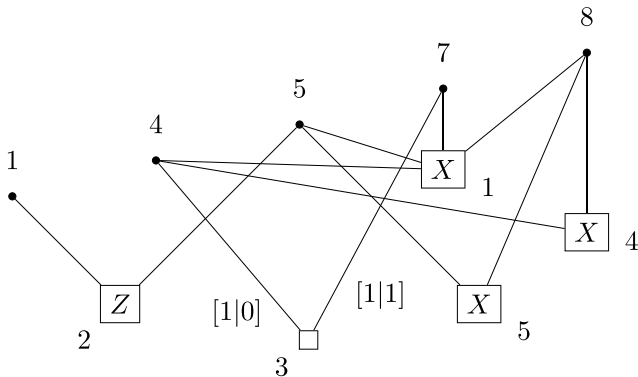
where the bottom right X check has two edges associated to it, as the X check has weight 4 on the logical operator $\bar{Z}_2$, see Ref. [40, Remark 4].

Finally, for $\bar{Y}_3$ we have

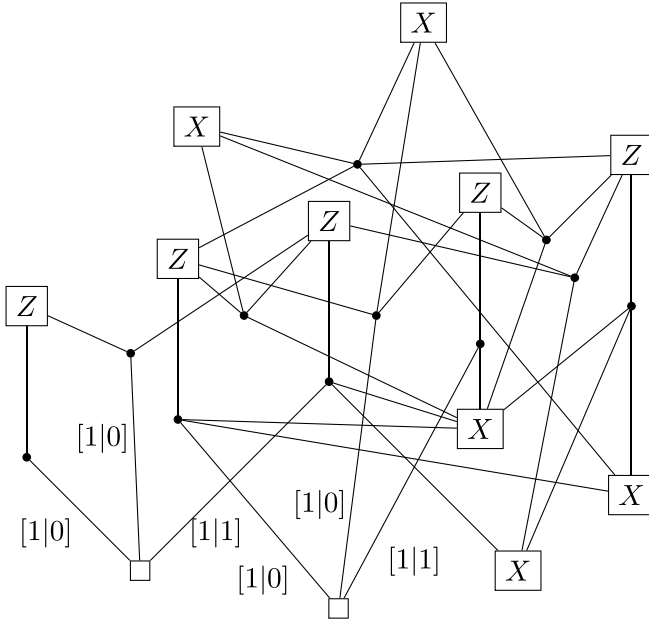
$$\left[\bar{Y}_3 \mid 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \mid 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 1 \right],$$

which is also incident to all checks in the $[[8, 3, 3]]$ code; all of the new checks added to branch $\bar{X}_1$ and $\bar{Z}_2$ commute with the action of $\bar{Y}_3$ on each qubit, so $\bar{Y}_3$ is left undeformed by the prior branching. The local code, consisting of checks which anticommute with the action of $\bar{Y}_3$ on each qubit, is then

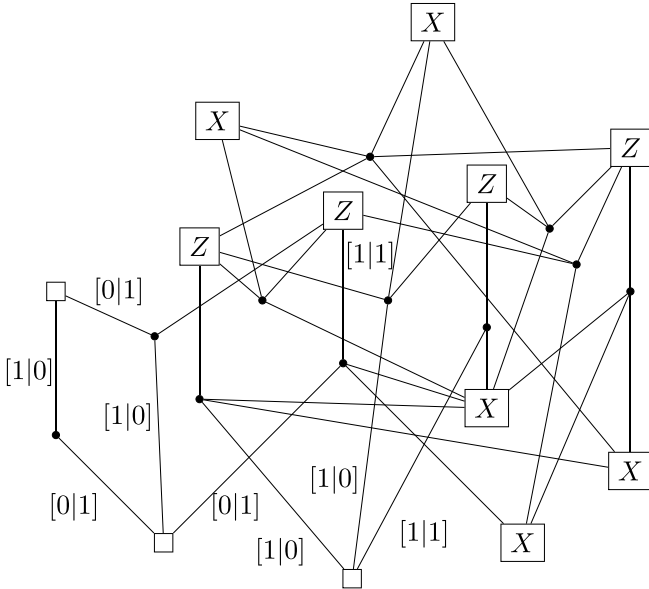
As $\bar{X}_1$ and $\bar{Z}_2$ have been branched away from the support of $\bar{Y}_3$, we measure $\bar{Y}_3$ directly. To do so, permute the actions of $\bar{Y}_3$ on data qubits to be in the Z -basis, applying H_1 and V_5 , where $V = HSH$ and $S = \sqrt{Z} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}$. The permuted local code is



where now $\bar{Y}_5$ acts as a product of Z Paulis. Then insert the measurement graph as before,



before finally permuting the bases back by applying H_1 and $V_5^\dagger$,



where, to avoid clutter, we have abused notation and kept the new check measuring $\bar{Y}_5$ a Z -check, but with a $[1|1]$ below to indicate the edge which is actually measuring Y .

All told, the scheme for measuring $\bar{X}_1, \bar{Z}_2, \bar{Y}_3$ simultaneously used 27 new qubits and 30 new checks. This overhead is largely because the original code is already dense, so performing LDPC code surgery largely just adds overhead without preserving any useful property. In practice, measuring the $[[8, 3, 3]]$ code could be done by adding three new checks, one for each of the logicals to be measured.

One relevant saving, compared to the upper bounds on cost, is that because all logical qubits are measured, and all cycles in the measurement graphs are gauge-fixed, the deformed code will have $k = 0$ and therefore $d = \infty$. As a consequence the measurement graphs do not need to be expanding to preserve code distance. All the other desiderata of Theorem 1 are satisfied. The same principle applies more generally: the more logical qubits are measured simultaneously on a quantum code, the fewer logical operators remain from the original code, so the less likely it is that cleaning them into the ancilla system will reduce the distance, see Refs. [40,41]. As a consequence, the higher the rate of measurement, the lower the expansion of measurement graphs is likely to need to be. Additionally, the X -checks required to gauge-fix cycles are small enough that decongestion would not be required to maintain any sparsity, with the maximum weight of these checks being 3.

APPENDIX C: BASIS OF LOGICAL OPERATORS FOR CSS CODES

Lemma 8. Consider a CSS code Q with check matrices H_X, H_Z and parameters $[[n, k, d]]$. We can find $\bar{X}$ logical operators $x_1, \dots, x_k \in \mathbb{F}_2^n$ and $\bar{Z}$ logical operators $z_1, \dots, z_k \in \mathbb{F}_2^n$ of this code, such that $|x_i \cap z_j| = \delta_{ij}$. Here $\cap$ denotes intersection as sets.

Proof. Let m_Z and m_X be the rank of H_Z and H_X , respectively. Let us row-reduce the matrix H_Z to the form $H_Z = [I_{m_Z}, R_Z]$, where I_{m_Z} denotes an identity matrix and R_Z is a $m_Z \times (n - m_Z)$ matrix, discarding any all-zero rows [79]. Let L_Z be a $k \times n$ matrix whose rows are a basis of Z logical operators of the code Q , and similar for L_X . Stack H_Z with L_Z to form the following $(m_Z + k) \times n$ matrix:

$$M_Z = \begin{bmatrix} I_{m_Z} & R_Z \\ L_Z & \end{bmatrix}.$$

Perform row-reduction on M_Z , but with the following rule: we can add any of the top m_Z rows (corresponding to rows of H_Z) to any of the bottom k rows, and we can add any of the bottom k rows to any of the bottom k rows. We are not allowed to add any of the bottom k rows to any of the top m_Z rows. Up to permutation of columns, we can reduce M_Z to have the form

$$M_Z = \begin{bmatrix} I_{m_Z} & R_Z \\ 0_{k, m_Z} & I_k & L'_Z \end{bmatrix},$$

where L'_Z is a $k \times (n - m_Z - k)$ matrix and $0_{k, m_Z}$ denote a $k \times m_Z$ matrix of zeros. Note that $n - m_Z - k = m_X$. From here on, we will fix the ordering of columns and no longer permute them. These last k rows are our Z logical operators z_i .

Now consider the matrix H_X . We know that the rows of H_X span the dual space of the row space of M_Z . Therefore, let us write down a basis of this dual space, which will be an equivalent set of X-stabilizer generators of Q . Let

$$H'_X = [R_X \quad I_{m_X}],$$

where R_X is a $m_X \times (n - m_X)$ matrix. We need to find R_X such that every row r of H'_X satisfies $M_Z r = 0$.

Let us write $R_Z = [R_{Z,1}, R_{Z,2}]$, where $R_{Z,1}$ is a $m_Z \times k$ matrix and $R_{Z,2}$ is a $m_Z \times m_X$ matrix. Then

$$M_Z = \begin{bmatrix} I_{m_Z} & R_{Z,1} & R_{Z,2} \\ 0_{k,m_Z} & I_k & L'_Z \end{bmatrix},$$

and we will take

$$H'_X = [(R_{Z,2})^\top + (R_{Z,1} L'_Z)^\top \quad (L'_Z)^\top \quad I_{m_X}].$$

One can check that $M_Z H'_X = 0$, which means the rows of H'_X form a valid basis of X-stabilizers. We can now stack H'_X with L_X and get

$$M_X = \begin{bmatrix} R_X & I_{m_X} \\ L_X & \end{bmatrix},$$

and we apply stabilizers to operators such that

$$M_X = \begin{bmatrix} R_X & I_{m_X} \\ L'_X & J \quad 0_{k,m_X} \end{bmatrix},$$

for some $k \times m_Z$ matrix L'_X and $k \times k$ matrix J . Note that the last k rows of M_X form a basis of X-logical operators of Q .

We claim that J must be full-rank. If not, there exists a vector x generated by the last k rows of M_X that has the form $x = [x', 0, \dots, 0]$, where x' is a length m_Z vector. Consequently, $H_Z x^\top = x'$, which contradicts the assumption that $H_Z M_X^\top = 0$. Therefore J must be full rank, we can reduce the last k rows (only adding these last k rows among themselves) of H_X to

$$[L''_X \quad I_k \quad 0_{k,m_X}].$$

These vectors are the X logical operators x_i of Q , which satisfy the desired intersection rules with the Z logical operators. ■

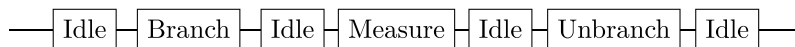
APPENDIX D: FAULT-TOLERANCE PROOFS

For all our proofs of fault-tolerance, we closely follow the approach in Ref. [40]. For this it is necessary to recall some definitions and conventions from the literature. We follow the general approach to fault-tolerance via repeated syndrome measurements [29], and the fault-distance is taken with respect to phenomenological noise; that is, there are faults on data qubits that can occur at any timestep, including initialization and measurement, and there are measurement faults on syndrome qubits that can occur at any timestep.

The choice of phenomenological noise is justified as follows. Consider performing bare ancilla syndrome extraction, introducing a qubit for each check, entangling the incident data qubits, then measuring out the check qubit. When the spacetime code is LDPC, the syndrome extraction circuit fault-distance is lower-bounded by d/c , where d is the phenomenological fault-distance of the spacetime code and c is a constant. See, for example, the discussion in Ref. [80, Sec. 5], or Ref. [81, Sec. 6.3]. The constant c is of vital importance in practice, and syndrome extraction circuits must be optimized to reduce it, but for the purposes of asymptotics it can be ignored. Moreover, as our parallel measurement scheme is generic, applicable to any stabilizer LDPC code, it is unfeasible to use circuit fault-distance directly, as we do not presume syndrome extraction circuits are defined in advance for the initial codes.

All our procedures begin at time t_0 followed by at least d rounds of syndrome measurement in the initial distance- d code Q . Then there is a code deformation step at time t_i with d rounds of syndrome measurement in the deformed code. When branching, we do not deform back immediately, but instead leave at least d rounds of syndrome measurement before future gauging logical measurements are applied. Then finally the code is unbranched at time t_o .

As the fault-tolerance of gauging logical measurements was proved in Ref. [40], we do not prove it here, but instead take as given that following branching there is an intermediate series of rounds of syndrome measurement such that upon exit of the gauging logical measurements procedure we then unbranch the code, ready for the next timestep in the logical circuit. That is, we have the circuit



where between each procedure and at each end there are d syndrome measurement rounds, labeled “Idle.” These “padding” steps are required for the proofs to prevent timelike errors entering the spacetime in which each code deformation procedure is performed; however, we expect they are not always necessary in general [82]. Indeed, we expect that the entire circuit above can be performed by initializing all new ancilla qubits, measuring all new checks for d syndrome measurement rounds and then measuring out all ancilla qubits, i.e., using only approximately d syndrome measurement rounds in total,



This more compact procedure would then be a single parallel gauging logical measurement, where the Z checks used to infer a given $\bar{Z}_i$ logical measurement include those which clean an initial v_i to its final location on the leaf of a branch. We do not prove that this is possible here, as such a proof would be cumbersome, and it does not affect the asymptotic time overhead.

We also follow Ref. [83] and use the convention that the initial and final rounds of measurement in the branching and unbranching procedures are perfect, which is justified by the “padding” as any fault that involves two code deformations, or both a code deformation and the initial or final boundary conditions, must have weight greater than d [40, Remark 29].

Definition D1 (Space and time faults). A space fault is a Pauli error operator that occurs on some qubit during the implementation of a code deformation. A time fault is a measurement error where the result of a measurement is reported incorrectly during the implementation of a code deformation. A general spacetime fault is a collection of space and time faults.

We consider state initialization faults to be time faults that are equivalent to space faults by decomposing them into a perfect initialization followed by an error operator.

Definition D2 (Detectors). A detector is a collection of state initializations and measurements that yield a deterministic result, in the sense that the product of the observed measurement results must be $+1$ independent of the individual measurement outcomes, if there are no faults in the procedure.

Remark D1. Some detectors have an empty collection of state initializations, i.e., they consist only of measurements: for example, repeated measurement of a stabilizer check constitutes a detector. Other detectors include both state initializations and measurements: for example, initialization of a CSS code in the $|0\rangle^{\otimes n}$ state, followed by measurement of Z -type checks. In our case, we have both

classes of detectors; the latter class arises from initializing new data qubits in the ancilla system then performing measurements upon them.

Definition D3 (Syndrome). The syndrome caused by a spacetime fault is defined to be the set of detectors that are violated in the presence of the fault. That is, the set of detectors which do not satisfy the constraint that the observed measurement results multiply to $+1$ in the presence of the fault.

We follow Ref. [40] in our convention for labeling time steps, whereby detectors and space errors occur at integer timesteps, but syndrome measurements and measurement errors occur at integer + half timesteps. Thus the first code deformation measurements are made at timestep $t_i + \frac{1}{2}$, and the final measurements before d rounds of padding are made at timestep $t_o + \frac{1}{2}$.

Check s_j is the j th check in the original code and $\tilde{s}_j$ the (possibly deformed) check from the original code in the new deformed code.

We consider the ancillary system to be an incidence structure, i.e., a chain complex, called C . While this incidence structure is not generally a cell complex, we nevertheless associate new Z checks to vertices, new data qubits to hyperedges and new X checks to faces; this is a labeling to match the similar proof in Ref. [40], where the incidence structure is strictly a cell complex.

We call a new Z check in the branching procedure an A_v check, corresponding to a vertex v . Similarly we have a B_p check on a face p in the X basis at time t .

Z_e and X_e act as Z and X Pauli operators on ancillary qubits associated with a hyperedge e . X_i and Z_i act as X and Z Pauli operators on a data qubit in the original code labeled i .

1. Proof of Lemma 3

Assume we are brute-force branching $\bar{Z}$ logicals. This does not lose generality, as the generalized branching from Lemma 4 is equivalent to branching $\bar{Z}$ logicals up to single-qubit Cliffords, using the same arguments as in Ref. [40]. Before time t_i we have the code Q , with distance d . All new data qubits are initialized in the $|+\rangle$ state at time t_i , and all checks are measured from time $t_i + \frac{1}{2}$ to $t_o + \frac{1}{2}$. All new X checks, in the absence of errors, deterministically return the $+1$ outcome. The new Z checks, however, return random outcomes due to the $|+\rangle$ input states. This scenario is slightly different from the scenario of gauging logical measurements, as there we know that the product of the new Z checks must yield the outcome of the logical measurement, in the absence of errors. Here we do not know the product of the outcomes. These outcomes dictate which Pauli frame [84] we are in, and so we can reinterpret consistent -1 outcomes as $+1$ consistent outcomes; in other words, there are detectors between each successive Z

measurement, but not a detector given by the first Z measurement and the initialization of the new qubits in the $|+\rangle$ state. Hence these outcomes are very similar to the state preparation of $|\bar{+}\rangle^{\otimes k}$ in a CSS code: the Z measurements induce a Pauli frame, which must then be consistent.

Throughout the discussion of the spacetime code, detectors (placed at integer timesteps) are always induced by at least one check. Therefore we use the notation that for a check C at time $t + \frac{1}{2}$ or $t - \frac{1}{2}$ a corresponding detector at time t which uses that check is labeled C^t .

Lemma D1 (Spacetime code for branching). The following form a generating set of the local detectors in the fault-tolerant branching procedure:

For $t < t_i$

- (1) s_j^t which is given by repeated s_j checks in the original code at times $t - \frac{1}{2}, t + \frac{1}{2}$.

For $t = t_i$

- (1) B_p^t which is given by the measurement of B_p at time $t_i + \frac{1}{2}$ together with the initialization of new qubits in the $|+\rangle$ state at time $t_i - \frac{1}{2}$.
- (2) $\tilde{s}_j^{t_i}$ which is given by the measurement of s_j , together with the initialization of new qubits in the $|+\rangle$ state at time $t_i - \frac{1}{2}$, and the measurement of $\tilde{s}_j$ at time $t_i + \frac{1}{2}$.

For $t > t_i$

- (1) A_v^t which is the repeated measurement of an A_v check in the deformed code at times $t - \frac{1}{2}$ and $t + \frac{1}{2}$.
- (2) B_p^t which is the same for a B_p check.
- (3) $\tilde{s}_j^t$ which is the same for an $\tilde{s}_j$ check.

Proof. Away from the initial and final steps of the code deformation, the brute-force branching procedure consists of repeatedly measuring the same checks. In this setting any local detector must contain a pair of measurements of the same check. Any detector formed by the measurement of the same check at times $(t, t+k)$ can be decomposed into detectors at time steps $(t, t+1), (t+1, t+2), \dots, (t+k-1, t+k)$. In other words, the time component of the spacetime code constitutes a repetition code [69].

During the code deformation there are also detectors involving the initialization of new qubits. Due to measurement of A_v terms during the code deformation, any detector involving the initialization of new qubits in the X basis must include a collection of new qubits that corresponds to a product of B_p and $\tilde{s}_j$ checks. Any such detector can be decomposed into those listed above which involve qubits at time t_i , combined with repeated measurement checks.

Local detectors that cross the deformation step at t_i must include a check s_j from the original code before or after the deformation, along with a deformed version $\tilde{s}_j$. Such a detector can be decomposed into repeated measurement detectors and one of the detectors involving new qubits listed above. ■

Spacetime faults can be organized according to the spacetime syndromes they create:

For $t < t_i$

- (1) An s_j measurement fault at time $t + \frac{1}{2}$ violates the detectors s_j^t and s_j^{t+1} .
- (2) A Pauli $Z_i(X_i)$ fault at time t violates the s_j^t detectors for all incident s_j which do not commute with $Z_i(X_i)$.

For $t = t_i$

- (1) A Pauli $Z_i(X_i)$ operator fault at time t violates the $\tilde{s}_j^t$ detectors for all $\tilde{s}_j$ which do not commute with $Z_i(X_i)$.
- (2) A Pauli Z_e fault at time t violates the B_p^t detectors for all p incident to e and the $\tilde{s}_j^t$ detectors for all $\tilde{s}_j$ checks which anticommute with Z_e .
- (3) A $|+\rangle_e$ initialization fault at time $t - \frac{1}{2}$ is equivalent to a Pauli Z_e operator fault at time t and so violates the same detectors.
- (4) An $\tilde{s}_j$ fault at time $t + \frac{1}{2}$ violates the detectors $\tilde{s}_j^t$ and $\tilde{s}_j^{t+1}$.

For $t > t_i$

- (1) A Pauli Z_i fault at time t violates the $\tilde{s}_j^t$ detectors for all $\tilde{s}_j$ which do not commute with Z_i .
- (2) A Pauli X_i fault at time t violates the $\tilde{s}_j^t$ detectors for all $\tilde{s}_j$ which do not commute with X_i , and the A_v^t detectors for all A_v which do not commute with X_i .
- (3) A Pauli X_e fault at time t violates the A_v^t detectors for all incident A_v checks.
- (4) A Pauli Z_e operator fault at time t violates the B_p^t detectors for all incident B_p checks, and $\tilde{s}_j$ detectors for all incident $\tilde{s}_j$ which anticommute with Z_e .
- (5) An $\tilde{s}_j$ measurement fault at time $t + \frac{1}{2}$ violates the detectors $\tilde{s}_j^t$ and $\tilde{s}_j^{t+1}$.
- (6) An A_v measurement fault at time $t + \frac{1}{2}$ violates the detectors A_v^t and A_v^{t+1} .
- (7) A B_p measurement fault at time $t + \frac{1}{2}$ violates the detector B_p^{t+1} .

Definition D4 (Spacetime logical fault). A spacetime logical fault is a collection of space and time faults that does not violate any detectors.

Definition D5 (Spacetime stabilizer). A spacetime stabilizer is a trivial spacetime logical fault in the sense that it

is a collection of space and time faults that does not violate any detectors and does not anticommute with any logical faults.

Lemma D2. The following form a generating set of local spacetime stabilizers:

For $t < t_i$

- (1) A stabilizer check operator s_j at time t .
- (2) A pair of X_i (or Z_i) faults at time $t, t+1$ together with measurement faults on all checks s_j that do not commute with X_i (or Z_i) at time $t + \frac{1}{2}$.

For $t = t_i$

- (1) A stabilizer check operator s_j or X_e at time t .
- (2) A pair of Z_i faults at time $t, t+1$ along with measurement faults on all checks $\tilde{s}_j$ and A_v that do not commute with Z_i at time $t+1$.
- (3) A pair of X_i faults at time $t, t+1$ along with measurement faults on all checks $\tilde{s}_j$ that do not commute with X_i at time $t+1$.
- (4) A pair of Z_e faults at times $t, t+1$, together with measurement faults on incident A_v checks and all incident $\tilde{s}_j$ checks that do not commute with Z_e at time $t + \frac{1}{2}$.
- (5) A $|+\rangle_e$ initialization fault at time $t - \frac{1}{2}$ together with a Z_e fault on that qubit at time t .
- (6) An X_e fault at time $t+1$ along with A_v measurement faults for all incident v at time $t + \frac{1}{2}$.

For $t > t_i$

- (1) A stabilizer check operator $\tilde{s}_j, A_v$, or B_p , at time t .
- (2) A pair of Z_i faults at time $t, t+1$ along with measurement faults on all checks $\tilde{s}_j$ that do not commute with Z_i at time $t+1$.
- (3) A pair of X_i faults at time $t, t+1$ along with measurement faults on all incident A_v checks, and all checks $\tilde{s}_j$ that do not commute with X_i at time $t+1$.
- (4) A pair of Z_e faults at time $t, t+1$ together with measurement faults on all incident B_p checks and all incident $\tilde{s}_j$ checks that do not commute with Z_e at time $t+1$.
- (5) A pair of X_e faults at time $t, t+1$, together with measurement faults on all incident A_v checks at time $t+1$.

Proof. First, recall from the organization of spacetime faults that all initialization errors are equivalent to Pauli operator errors. Then, any nonempty local spacetime stabilizer must involve a Pauli operator error, as otherwise the stabilizer would have to include all repeated measurement errors of some check.

A nonempty spacetime stabilizer containing a Pauli operator must either be a space stabilizer or contain a Pauli

operator at another time. The product of the Pauli operators from all time steps involved must itself be a space stabilizer. Any local spacetime stabilizer of this form can be constructed from a product of the spacetime stabilizers introduced above by first reconstructing the operators at the earliest time step at the cost of creating matching operators at the next time step, and so on until the final time step, where the product of the operators must now become trivial. This leaves a local spacetime stabilizer with only measurement errors, which must also be trivial. Hence, the original fault pattern is a product of the introduced spacetime stabilizers as claimed. ■

Definition D6 (Spacetime fault-distance). The spacetime fault-distance is the weight, counted in terms of single site Pauli errors and single measurement errors, of the minimal collection of faults that does not violate any detectors and is not a spacetime stabilizer.

Lemma D3 (Time fault-distance). The fault-distance for measurement and initialization errors is at least $(t_o - t_i)$.

Proof. All pure measurement logical faults must start at the code deformation step at time t_i and end at time t_o . Otherwise, a measurement fault at time step t must be followed and preceded by another measurement fault on the same type of check at time steps $t-1, t+1$. Thus a measurement and initialization logical fault must have distance at least $(t_o - t_i)$. ■

Lemma D4 (Decoupling of space and time faults). Any nontrivial spacetime logical fault is equivalent to the product of a space logical fault and a time logical fault, up to multiplication with spacetime stabilizers.

Proof. Let F be an arbitrary spacetime fault. The space component of F can be cleaned into a chosen single timestep $t_i \leq t \leq t_o$ by multiplication with the spacetime stabilizers listed from Lemma D2 that involve like Pauli operator faults at time steps $t, t+1$ along with the appropriate measurement faults. Therefore we can clean all the space component of F to time t_i .

In the cleaned spacetime logical fault all measurement errors must occur in the time steps $t_i \leq t \leq t_o$ as measurement errors outside this time window must propagate to the initial or final time boundary, which has no measurement errors by assumption. The measurement faults now form strings that propagate through time from $t_i + \frac{1}{2}$ to $t_o + \frac{1}{2}$, terminating at an A_v or B_p measurement fault, and so are time logical faults forming codewords in the repetition code timelike component of the spacetime code. ■

Lemma D5 (Spacetime fault-distance). Brute-force branching is a fault-tolerant operation with fault-distance d , assuming we use at least d rounds of stabilizer measurements.

Proof. Let F be any spacetime logical fault that is not equivalent to a spacelike logical fault. Then F must have support on all time steps $t + \frac{1}{2}$ such that $t_i \leq t \leq t_o$, and so if $t_o - t_i \geq d$ then their weight is bounded below by d .

Now let F be a spacetime logical fault that is equivalent to a spacelike logical fault. We will clean F to become a spacelike logical fault using the spacetime stabilizers. Observe that we can always clean F to become a spacelike logical fault using the spacetime stabilizer generators from Lemma D2 without increasing the weight, as each stabilizer preserves the parity of space and initialization faults at a fixed location. That cleaned spacelike logical fault must have weight at least d , by Lemma 2. Therefore the weight of F must be at least d . ■

That unbranching after a gauging logical measurement is also a fault-tolerant operation can be deduced in a similar manner, where now every new data qubit is measured out in the X basis at time $t = t_o$. Unbranching only uses one stabilizer check round of measurements, but we assume d rounds of padding before and after, as previously. This gives a similar spacetime code.

Lemma D6 (Spacetime code for unbranching). The following form a generating set of the local detectors in the fault-tolerant unbranching procedure.

For $t < t_o$

- (1) A_v^t which is the repeated measurement of an A_v check in the deformed code at times $t - \frac{1}{2}$ and $t + \frac{1}{2}$.
- (2) B_p^t which is the same for a B_p check.
- (3) $\tilde{s}_j^t$ which is the same for an $\tilde{s}_j$ check.

For $t = t_o$

- (1) B_p^t which is the measurement of B_p at time $t_o - \frac{1}{2}$ along with the measurement of X_e on each incident hyperedge data qubit at time $t_o + \frac{1}{2}$.
- (2) $\tilde{s}_j^{t_o}$ which is the measurement of $\tilde{s}_j$ at time $t_o - \frac{1}{2}$ along with the measurement of X_e on each incident hyperedge data qubit, and s_j at time $t_o + \frac{1}{2}$.

For $t > t_o$

- (1) s_j^t which is given by the collection of repeated s_j checks in the original code at times $t - \frac{1}{2}, t - \frac{1}{2}$.

Observe that the spacetime detectors for $t < t_o$ are the same as for $t > t_i$ when branching. The same is true for spacetime faults and spacetime stabilizers for $t < t_o$. Similarly, the spacetime detectors, faults, and stabilizers for $t > t_o$ are the same as for $t < t_i$ when branching.

The spacetime faults at $t = t_o$, along with the spacetime syndromes they yield, are given by

- (1) An X_i (or Z_i) operator fault at time t violates the $\tilde{s}_j^t$ detectors for all $\tilde{s}_j$ which do not commute with X_i (or Z_i).
- (2) A Z_e operator fault on a new data qubit at time t violates the $\tilde{s}_j^t$ and B_p^t detectors for all $\tilde{s}_j$ and B_p which do not commute with Z .
- (3) An X_e -measurement read-out fault at time $t + \frac{1}{2}$ is equivalent to a Z_e operator fault at time t and violates the same detectors.
- (4) An s_j measurement fault at time $t + \frac{1}{2}$ violates the detectors $\tilde{s}_j^t$ and s_j^{t+1} .
- (5) A B_p measurement fault at time $t + \frac{1}{2}$ violates the detector B_p^{t-1} .

The spacetime stabilizers at $t = t_o$ have the generating set

- (1) A stabilizer check operator $\tilde{s}_j, A_v$ or B_p at time t .
- (2) A pair of X_i (Z_i) operator faults at time $t, t + 1$, together with measurement faults on all checks s_j which do not commute with X_i (Z_i).
- (3) A Pauli Z_e fault on a hyperedge data qubit at time t , along with an X_e -measurement read-out fault at time $t + \frac{1}{2}$.
- (4) A Pauli X_e fault on a hyperedge data qubit at time t .
- (5) A Pauli X_e error on a hyperedge data qubit at time $t - 1$ along with measurement errors on incident A_v checks at time $t - \frac{1}{2}$.

Together with the spacetime detectors, faults, and stabilizers before and after the unbranching, one can then use the same arguments as in Lemma D5 to conclude that unbranching has fault-distance d . Because of the “padding,” we need only measure out the new data qubits and pre-existing stabilizers once. We argue that, as with branching and gauging logical measurements, this padding is typically unnecessary in practice.

APPENDIX E: SIMULTANEOUS BRANCHING

Consider the case where we wish to perform $\bar{Z}$ and $\bar{X}$ measurements on t different logical qubits in parallel, but the representatives we have chosen are overlapping. In this case, we can still brute-force branch the $\bar{Z}$ logicals, say, but the $\bar{X}$ logicals which overlapped with them will be deformed to have new support on the branch [47, Appendix F].

In particular, let $\{v_1, v_2, \dots, v_{t_Z}\}$ be the $\bar{Z}$ logicals, and $V = \bigcup_i v_i$ be the union of data qubit support of these logicals.

Lemma E1. Let each $\bar{X}$ logical w_j to be measured commute with each $\bar{Z}$ logical contained in V . Then each $\bar{X}$ logical can be deformed such that it picks up support in the first layer of the Z -type tree.

Proof. There are two branches attached to V , one with support on u_1 and one on u_2 , with u_1 and u_2 being the unions of support for the first $\lceil t/2 \rceil$ logicals and the last $\lfloor t/2 \rfloor$, respectively. We have $u_1 \subset V$ and $u_2 \subset V$. For the first branch, let ∂_1 be H_X restricted to u_1 . Then the first layer of the first branch is a Z layer, with a code from data qubits to Z checks given by $\partial_1^\top$. An overlapping w_j will now anticommute with the Z checks in the first layer in correspondence with $w_j \cap V$. As w_j commutes with each logical in V , $w_j \cap u_1 \in \ker(\partial_1)^\perp$, and hence $w_j \cap u_1$ is in the column space of $\partial_1^\top$. Therefore there is some set of new data qubits in the first layer such that extending w_j to these data qubits, i.e., applying X to those data qubits, turns off the Z checks in bijection with $w_j \cap u_1$. As the next layer is an X layer, there are no further anticommuting checks.

Repeating this procedure for the other branch gives a set of qubits in the first layer of that branch as well. ■

We can see that, in the best case, w_j only overlaps with u_1 and not u_2 , so only new qubits in one branch are added to the support of w_j ; in the worst case w_j could have substantial overlap with both subsets and hence needs support on both branches. In the best case, the number of new qubits required is equal to the number of checks c which are applied to clean w_j from V ; in the worst case it could be $2c$.

Also observe that strictly speaking we do not need w_j to commute with every $\bar{Z}$ logical in V : only each $\bar{Z}$ logical in u_1 and u_2 . Therefore if the first layer of branching results in codes ∂_1 and ∂_2 , which remove $\bar{Z}$ logicals which w_j anticommuted with then the deformation to the first layer still applies.

Lemma E2. Let each $\bar{X}$ logical w_j to be measured commute with each $v_1, v_2, \dots, v_t$. Then each $\bar{X}$ logical can be deformed such that it picks up support in the Z -type tree.

Proof. In this case there are some anticommuting $\bar{Z}$ logicals contained in V , but w_j still commutes with the logicals to be measured. When w_j commutes with every $\bar{Z}$ logical in u_1 and u_2 we can merely add support on the first layer as before. When it does not, we can add support to the second layer of the first branch in bijection with $w_j \cap u_1$ to satisfy the Z checks in the first layer. This then anticommutes with the Z checks in the third layer. Continue until we have support extending through the tree, until at some layer every $\bar{Z}$ logical which anticommuted with w_j has been removed. In the worst case we add the remaining support to w_j on the last Z layer, which is the penultimate layer of the tree, as the last layer is an X layer with only each v_i on leaves, and $[w_j, v_i] = 0$.

Then repeat this procedure for the other branch with support on u_2 . ■

In the worst case the $\bar{X}$ logical w_j can pick up support on $\mathcal{O}(t \omega \log t)$ new qubits. This is because w_j can pick up support on each level of the branching tree, with up to weight $\mathcal{O}(t \omega)$ on each level, and there are $\mathcal{O}(\log t)$ levels.

We can then branch the deformed $\bar{X}$ logicals to measure them separately, so in one logical Pauli measurement timestep we branch $\bar{Z}$ logicals and $\bar{X}$ logicals and then measure with gauging logical measurements. As the deformed $\bar{X}$ logicals have no support on the leaves, this X -type branching leaves the $\bar{Z}$ measurements unaffected. Because the procedure may increase the weights of $\bar{X}$ to be measured by $\mathcal{O}(t \omega \log t)$, and the procedure gives no clear method to measure $\bar{Y}$ logical operators in parallel, we focus on the method of choosing a suitable basis and representatives for our logical operators which avoid overlap, thereby avoiding these problems.

In finite-size practical cases, the fact that the representatives to be measured could be much smaller initially than those given by Lemma 8 could lead to branching with overlap being more efficient.

We argue that, as in Appendix D, the branches can be initialized simultaneously, at the same time as the data qubits required for gauging measurements, and the whole procedure performed with d timesteps in total while maintaining fault-distance d , but we do not prove this here.

-
- [1] Google Quantum AI, Suppressing quantum errors by scaling a surface code logical qubit, *Nature* **614**, 676 (2023).
 - [2] Google Quantum AI and Collaborators, Quantum error correction below the surface code threshold, *Nature* **638**, 920 (2024).
 - [3] D. Bluvstein *et al.*, Logical quantum processor based on reconfigurable atom arrays, *Nature* **626**, 58 (2024).
 - [4] Pedro Sales Rodriguez *et al.*, Experimental demonstration of logical magic state distillation, *Nature* **645**, 620 (2025).
 - [5] N. Lacroix *et al.*, Scaling and logic in the color code on a superconducting quantum processor, *Nature* **645**, 614 (2025).
 - [6] N. P. Breuckmann and J. N. Eberhardt, Quantum low-density parity-check codes, *PRX Quantum* **2**, 040101 (2021).
 - [7] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, High-threshold and low-overhead fault-tolerant quantum memory, *Nature* **627**, 778 (2024).
 - [8] Q. Xu, J. P. B. Ataiades, C. A. Pattison, N. Raveendran, D. Bluvstein, J. Wurtz, B. Vasić, M. D. Lukin, L. Jiang, and H. Zhou, Constant-overhead fault-tolerant quantum computation with reconfigurable atom arrays, *Nat. Phys.* **20**, 1084 (2024).
 - [9] A. Yu. Kitaev, Fault-tolerant quantum computation by anyons, *Ann. Phys.* **303**, 2 (2003).
 - [10] D. Horsman, A. G. Fowler, S. Devitt, and R. V. Meter, Surface code quantum computing by lattice surgery, *New J. Phys.* **14**, 123011 (2012).
 - [11] J. E. Moussa, Transversal Clifford gates on folded surface codes, *Phys. Rev. A* **94**, 042316 (2016).

- [12] C. Gidney, N. Shutty, and C. Jones, Magic state cultivation: Growing T states as cheap as CNOT gates, [arXiv:2409.17595](#).
- [13] Markus Grassl and Martin Roetteler, in *2013 IEEE International Symposium on Information Theory* (IEEE, Istanbul, Turkey, 2013), pp. 534–538.
- [14] N. P. Breuckmann and S. Burton, Fold-transversal Clifford gates for quantum codes, [Quantum 8, 1372 \(2024\)](#).
- [15] A. O. Quintavalle, P. Webster, and M. Vasmer, Partitioning qubits in hypergraph product codes to implement logical gates, [Quantum 7, 1153 \(2023\)](#).
- [16] J. N. Eberhardt and V. Steffan, Logical operators and fold-transversal gates of bivariate bicycle codes, [IEEE Trans. Inf. Theory 71, 1140 \(2025\)](#).
- [17] Hasan Sayginel, Stergios Koutsoumpas, Mark Webster, Abhishek Rajput, and Dan E. Browne, Fault-tolerant logical Clifford gates from code automorphisms, [PRX Quantum 6, 030343 \(2025\)](#).
- [18] A. J. Malcolm, A. N. Glauddell, P. Fuentes, D. Chandra, A. Schotte, C. DeLisle, R. Haenel, A. Ebrahimi, J. Roffe, A. O. Quintavalle, S. J. Beale, N. R. Lee-Hone, and S. Simmons, Computing efficiently in QLDPC codes, [arXiv:2502.07150](#).
- [19] Guanyu Zhu, Shehryar Sikander, Elia Portnoy, Andrew W. Cross, and Benjamin J. Brown, Non-Clifford and parallelizable fault-tolerant logical gates on constant and almost-constant rate homological quantum low-density parity-check codes via higher symmetries, [PRX Quantum 6, 040361 \(2025\)](#).
- [20] Thomas R. Scruby, Arthur Pesah, and Mark Webster, Quantum rainbow codes, [arXiv:2408.13130](#).
- [21] Louis Golowich and Ting-Chun Lin, in *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC '25* (Association for Computing Machinery, New York, NY, 2025), pp. 689–696.
- [22] Ting-Chun Lin, Transversal non-Clifford gates for quantum LDPC codes on sheaves, [arXiv:2410.14631](#).
- [23] Po-Shen Hsin, Ryohei Kobayashi, and Guanyu Zhu, Classifying logical gates in quantum codes via cohomology operations and symmetry, [arXiv:2411.15848](#).
- [24] Nikolas P. Breuckmann, Margarita Davydova, Jens N. Eberhardt, and Nathanan Tantivasadakarn, Cups and gates I: Cohomology invariants and logical quantum operations, [Commun. Math. Phys. 407, 86 \(2026\)](#).
- [25] Guanyu Zhu, A topological theory for qLDPC: Non-Clifford gates and magic state fountain on homological product codes with constant rate and beyond the $N^{1/3}$ distance barrier, [arxiv:2501.19375](#).
- [26] S. Huang, T. Jochym-O'Connor, and T. J. Yoder, Homomorphic logical measurements, [PRX Quantum 4, 030301 \(2023\)](#).
- [27] Q. Xu, H. Zhou, G. Zheng, D. Bluvstein, J. P. B. Ataiades, M. D. Lukin, and L. Jiang, Fast and parallelizable logical computation with homological product codes, [Phys. Rev. X 15, 021065 \(2025\)](#).
- [28] L. Z. Cohen, I. H. Kim, S. D. Bartlett, and B. J. Brown, Low-overhead fault-tolerant quantum computing using long-range connectivity, [Sci. Adv. 8, eabn1717 \(2022\)](#).
- [29] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, Topological quantum memory, [J. Math. Phys. 43, 4452 \(2002\)](#).
- [30] H. Zhou, Chen Zhao, Madelyn Cain, Dolev Bluvstein, Nishad Maskara, Casey Duckering, Hong-Ye Hu, Sheng-Tao Wang, Aleksander Kubica, and Mikhail D. Lukin, Low-overhead transversal fault tolerance for universal quantum computation, [Nature 646, 303 \(2025\)](#).
- [31] B. Eastin and E. Knill, Restrictions on transversal encoded quantum gate sets, [Phys. Rev. Lett. 102, 110502 \(2009\)](#).
- [32] See also introduction of Ref. [85] for a discussion on the challenge of performing addressable logic with transversal gates.
- [33] S. Bravyi, G. Smith, and J. A. Smolin, Trading classical and quantum computational resources, [Phys. Rev. X 6, 021043 \(2016\)](#).
- [34] In Ref. [86], for instance, a circuit made of Clifford and T gates is compiled into Pauli measurements with ancillary T magic states, rendering the application of logical Clifford gates unnecessary, in principle.
- [35] Another subtle difference between generalized code surgery and homomorphic measurement is that surgery has more flexible control over which set of logical qubits one wishes to address, while homomorphic measurement is, prior to this work, more parallel than surgery.
- [36] C. Vuillot, L. Lao, B. Criger, C. G. Almudéver, K. Bertels, and B. M. Terhal, Code deformation and lattice surgery are gauge fixing, [New J. Phys. 21, 033028 \(2019\)](#).
- [37] M. B. Hastings, Weight reduction for quantum codes, [arXiv:1611.03790](#).
- [38] M. B. Hastings, On quantum weight reduction, [arXiv:2102.10030](#).
- [39] E. Sabo, L. G. Gunderman, B. Ide, M. Vasmer, and G. Dauphinais, Weight-reduced stabilizer codes with lower overhead, [PRX Quantum 5, 040302 \(2024\)](#).
- [40] Dominic J. Williamson and Theodore J. Yoder, Low-overhead fault-tolerant quantum computation by gauging logical operators, [Nat. Phys. 22, 598 \(2026\)](#).
- [41] B. Ide, M. G. Gowda, P. J. Nadkarni, and G. Dauphinais, Fault-tolerant logical measurements via homological measurement, [Phys. Rev. X 15, 021088 \(2025\)](#).
- [42] A. Cowtan and S. Burton, CSS code surgery as a universal construction, [Quantum 8, 1344 \(2024\)](#).
- [43] A. Cowtan, SSIP: Automated surgery with quantum LDPC codes, [arXiv:2407.09423](#).
- [44] A. Cross, Z. He, P. Rall, and T. J. Yoder, Improved QLDPC surgery: Logical measurements and bridging codes, [arXiv:2407.18393](#).
- [45] M. Freedman and M. Hastings, Building manifolds from quantum codes, [Geom. Funct. Anal. 31, 855 \(2021\)](#).
- [46] Esha Swaroop, Tomas Jochym-O'Connor, and Theodore J. Yoder, Universal adapters between quantum low-density parity check codes, [PRX Quantum 7, 010324 \(2026\)](#).
- [47] Guo Zhang and Ying Li, Time-efficient logical operations on quantum low-density parity check codes, [Phys. Rev. Lett. 134, 070602 \(2025\)](#).
- [48] Note that even if $\omega \in \mathcal{O}(n)$, d will be asymptotically much larger than $\log^3 \omega$ unless d is exponentially smaller than n . Thus when taking the same logical representatives as input, our scheme is asymptotically more efficient than Zhang and Li in most reasonable cases, with hyperbolic surface codes a notable exception [87].

- [49] C. Chamberland and E. T. Campbell, Universal quantum computing with twist-free and temporally encoded lattice surgery, *PRX Quantum* **3**, 010331 (2022).
- [50] Daniel Gottesman, Stabilizer codes and quantum error correction. Ph.D. thesis, California Institute of Technology, 1997.
- [51] Benjamin Ide, Manoj G. Gowda, Priya J. Nadkarni, and Guillaume Dauphinais, Fault-tolerant logical measurements via homological measurement, *Phys. Rev. X* **15**, 021088 (2025).
- [52] A. Hamma, P. Zanardi, and X.-G. Wen, String and membrane condensation on three-dimensional lattices, *Phys. Rev. B* **72**, 035307 (2005).
- [53] P. Panteleev and G. Kalachev, in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2022* (Association for Computing Machinery, New York, NY, 2022), pp. 375–388.
- [54] Z. He, A. Cowtan, D. J. Williamson, and T. J. Yoder, Extractors: QLDPC architectures for efficient Pauli-based computation, [arXiv:2503.10390](https://arxiv.org/abs/2503.10390).
- [55] T. J. Yoder, E. Schoute, P. Rall, E. Pritchett, J. M. Gambetta, A. W. Cross, M. Carroll, and M. E. Beverland, Tour de gross: A modular quantum computer based on bivariate bicycle codes, [arXiv:2506.03094](https://arxiv.org/abs/2506.03094).
- [56] Terminology borrowed from Ref. [88].
- [57] J.-P. Tillich and G. Zémor, Quantum LDPC codes with positive rate and minimum distance proportional to the square root of the blocklength, *IEEE Trans. Inf. Theory* **60**, 1193 (2014).
- [58] Throughout this paper the branching ancillary system has 1-to-1 connections from the first layer of a branch to the appropriate subset of logicals being branched. In Ref. [47] this is left as a variable, so one can choose this connectivity.
- [59] S. Bravyi and B. Terhal, A no-go theorem for a two-dimensional self-correcting quantum memory based on stabilizer codes, *New J. Phys.* **11**, 043029 (2009).
- [60] The first version of this theorem required the logical operators to be irreducible but was later generalized to remove this requirement.
- [61] Yu Shi, Ashlesha Patil, and Saikat Guha, Stabilizer entanglement distillation and efficient fault-tolerant encoders, *PRX Quantum* **6**, 010339 (2025).
- [62] David Aasen, Jeongwan Haah, Matthew B. Hastings, and Zhenghan Wang, Geometrically enhanced topological quantum codes, [arXiv:2505.10403](https://arxiv.org/abs/2505.10403).
- [63] Alexander Vardy, The intractability of computing the minimum distance of a code, *IEEE Trans. Inf. Theory* **43**, 1757 (1997).
- [64] I. Dumer, D. Micciancio, and M. Sudan, Hardness of approximating the minimum distance of a linear code, *IEEE Trans. Inf. Theory* **49**, 22 (2003).
- [65] The terminology comes from the fact that Y basis measurements can be performed using “twists” in surface code patches. We are not using surface codes, but use the same terminology for consistency.
- [66] I. H. Kim, Y.-H. Liu, S. Pallister, W. Pol, S. Roberts, and E. Lee, Fault-tolerant resource estimate for quantum chemical simulations: Case study on Li-ion battery electrolyte molecules, *Phys. Rev. Res.* **4**, 023019 (2022).
- [67] Min-Hsiu Hsieh, Xingjian Li, and Ting-Chun Lin, Simplified quantum weight reduction with optimal bounds, [arXiv:2510.09601](https://arxiv.org/abs/2510.09601).
- [68] Guo Zheng, Liang Jiang, and Qian Xu, High-rate surgery: Towards constant-overhead logical operations, [arXiv:2510.08523](https://arxiv.org/abs/2510.08523).
- [69] Timo Hillmann, Guillaume Dauphinais, Ilan Tzitrin, and Michael Vasmer, Single-shot and measurement-based quantum error correction via fault complexes, *Phys. Rev. A* **112**, L040401 (2025).
- [70] Nouédy Baspin, Lucas Berent, and Lawrence Z. Cohen, Fast surgery for quantum LDPC codes, [arXiv:2510.04521](https://arxiv.org/abs/2510.04521).
- [71] Alexander Cowtan, Zhiyang He, Dominic J. Williamson, and Theodore J. Yoder, Fast and fault-tolerant logical measurements: Auxiliary hypergraphs and transversal surgery, [arXiv:2510.14895](https://arxiv.org/abs/2510.14895).
- [72] TikZit, <https://tikzit.github.io/index.html>.
- [73] This notation comes from the fact that the hypergraph product is an internal hom $[-, -]$ in the compact closed category of finite-dimensional bounded chain complexes over $\mathbb{F}_2$ [89].
- [74] B. Audoux and A. Couvreur, On tensor products of CSS codes, *Ann. Inst. Henri Poincaré Comb. Phys. Interact.* **6**, 239 (2019).
- [75] C. A. Weibel, *An Introduction to Homological Algebra*, Cambridge Studies in Advanced Mathematics (Cambridge University Press, 1994).
- [76] Alexander Cowtan, Homology, Hopf algebras and quantum code surgery, [arXiv:2508.01496](https://arxiv.org/abs/2508.01496).
- [77] Indeed, $I \cap J$ does not need to be zero, which allows for branching the same logical operator to multiple different leaves.
- [78] Arthur Pesah, Austin K. Daniel, Ilan Tzitrin, and Michael Vasmer, Fault-tolerant transformations of spacetime codes, [arXiv:2509.09603](https://arxiv.org/abs/2509.09603).
- [79] This row echelon basis of stabilizers was studied in Ref. [90] in the context of quantum code testing.
- [80] Daniel Gottesman, Fault-tolerant quantum computation with constant overhead, *Quantum Inf. Comput.* **14**, 1338 (2014).
- [81] Zhiyang He, Quynh T. Nguyen, and Christopher A. Pattison, Composable quantum fault-tolerance, [arXiv:2508.08246](https://arxiv.org/abs/2508.08246).
- [82] Regarding the idling steps at either end of the circuit, it is common in the study of fault-tolerant protocols to insert rounds of perfect syndrome extraction immediately before and after the protocol. However, this leads to a well-known composition problem [83]. When fault-tolerant protocols proved in this manner are performed sequentially they are no longer generally fault tolerant; cases where this can occur are described in Ref. [71]. Inserting d rounds of idling ensures that any nontrivial logical spacetime fault which extends from one protocol to another must have weight at least d , and so composition is preserved.
- [83] M. E. Beverland, S. Huang, and V. Kliuchnikov, Fault tolerance of stabilizer channels, [arXiv:2401.12017](https://arxiv.org/abs/2401.12017).
- [84] E. Knill, Quantum computing with realistically noisy devices, *Nature* **434**, 39 (2005).

- [85] Z. He, V. Vaikuntanathan, A. Wills, and R. Y. Zhang, Quantum codes with addressable and transversal non-Clifford gates, [arXiv:2502.01864](#).
- [86] D. Litinski, A game of surface codes: Large-scale quantum computing with lattice surgery, [Quantum](#) **3**, 128 (2019).
- [87] Nikolas P. Breuckmann, Christophe Vuillot, Earl Campbell, Anirudh Krishna, and Barbara M. Terhal, Hyperbolic and semi-hyperbolic surface codes for quantum storage, [Quantum Sci. Technol.](#) **2**, 035007 (2017).
- [88] T. Carette, D. Horsman, and S. Perdrix, in *44th International Symposium on Mathematical Foundations of Computer Science* (MFCS, 2019), *Leibniz International Proceedings in Informatics (LIPIcs)*, Vol. 138, pp. 55:1–55:15, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2019).
- [89] S. MacLane, *Categories for the Working Mathematician*, Graduate Texts in Mathematics, Vol. 5 (Springer, 1997), 2nd ed.
- [90] A. Cross, Z. He, A. Natarajan, M. Szegedy, and G. Zhu, Quantum locally testable code with constant soundness, [Quantum](#) **8**, 1501 (2024).