Quantum Routing and Entanglement Dynamics Through Bottlenecks

Dhruv Devulapalli,[1, 2, *] Chao Yin,[3, *] Andrew Y. Guo,[4] Eddie Schoute,[5, 6]
Andrew M. Childs,[1, 7, 8] Alexey V. Gorshkov,[1, 2] and Andrew Lucas[3]

[1]*Joint Center for Quantum Information and Computer Science,*
NIST/University of Maryland, College Park, MD 20742, USA
[2]*Joint Quantum Institute, NIST/University of Maryland, College Park, MD 20742, USA*
[3]*Department of Physics and Center for Theory of Quantum Matter,*
University of Colorado, Boulder, CO 80309, USA
[4]*Quantinuum, Broomfield, CO, 80021, USA*
[5]*Computer, Computational, and Statistical Sciences Division,*
Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[6]*IBM Quantum, MIT-IBM Watson AI Lab, Cambridge, MA 02142, USA*
[7]*Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA*
[8]*Department of Computer Science, University of Maryland, College Park, MD 20742, USA*

To implement arbitrary quantum circuits in architectures with restricted interactions, one may effectively simulate all-to-all connectivity by routing quantum information. We consider the entanglement dynamics and routing between two regions only connected through an intermediate "bottleneck" region with few qubits. In such systems, where the entanglement rate is restricted by a vertex boundary rather than an edge boundary of the underlying interaction graph, existing results such as the small incremental entangling theorem give only a trivial constant lower bound on the routing time (the minimum time to perform an arbitrary permutation). We significantly improve the lower bound on the routing time in systems with a vertex bottleneck. Specifically, for any system with two regions L, R with N_L, N_R qubits, respectively, coupled only through an intermediate region C with N_C qubits, for any $\delta > 0$ we show a lower bound of $\Omega(N_R^{1-\delta}/\sqrt{N_L N_C})$ on the Hamiltonian quantum routing time when using piecewise time-independent Hamiltonians, or time-dependent Hamiltonians subject to a smoothness condition. We also prove an upper bound on the average amount of bipartite entanglement between L and C, R that can be generated in time t by such architecture-respecting Hamiltonians in systems constrained by vertex bottlenecks, improving the scaling in the system size from $O(N_L t)$ to $O(\sqrt{N_L} t)$. As a special case, when applied to the star graph (i.e., one vertex connected to N leaves), we obtain an $\Omega(\sqrt{N^{1-\delta}})$ lower bound on the routing time and on the time to prepare $N/2$ Bell pairs between the vertices. We also show that, in systems of free particles, we can route optimally on the star graph in time $\Theta(\sqrt{N})$ using Hamiltonian quantum routing, obtaining a speed-up over gate-based routing, which takes time $\Theta(N)$.

1. INTRODUCTION

The promise of wide-ranging applications of quantum computing has sparked interest in developing scalable quantum architectures. Realistic quantum architectures have constrained interactions, often with fixed connectivity [1, 2]. This connectivity is usually specified by a graph G, with vertices V representing qubits, and edges E representing pairs of qubits between which interactions are permitted. On the other hand, quantum algorithms, quantum error correction schemes, and other quantum information-processing protocols often require interactions between arbitrary pairs of qubits. To implement these interactions, quantum compilers effectively simulate all-to-all connectivity by moving quantum information, at the cost of an overhead in the running time. This task, known as quantum routing, aims to make general quantum algorithms feasible on devices with connectivity constraints by efficiently implementing any permutation

of qubits on a given architecture graph [3]. Quantum routing can be performed with swap gates using classical techniques [3–7], or employing more general quantum operations such as continuous-time Hamiltonian evolution [8, 9] or unitary gates [9], possibly assisted by measurement and fast classical feedback [10–12]. Recent work has shown that the worst-case depth overhead of implementing arbitrary quantum circuits while respecting connectivity constraints is proportional to the circuit depth of worst-case routing [13], highlighting the importance of lower bounds and fast protocols for this task.

Routing is closely related to the task of entanglement distribution, since by routing entangled qubits, we can distribute entanglement. The rate at which entanglement can be generated between parts of a system constrains our ability to perform many quantum information processing tasks, as entanglement is a critical resource for quantum algorithms [14–18], quantum error correction [19–21], quantum communication [22–24], and quantum sensing [25–27]. The fundamental importance of entanglement and its application to a wide range of quantum information tasks has raised many questions and led to a large number of investigations into the dynamics of entanglement [28–34].

FIG. 1. The star graph on 9 vertices, S_9.

FIG. 2. A tripartition with vertex bottleneck C. Parts L, C, R contain N_L, N_C, N_R qubits, respectively. By the tripartite connectivity constraint, there are no edges between qubits in L and qubits in R.
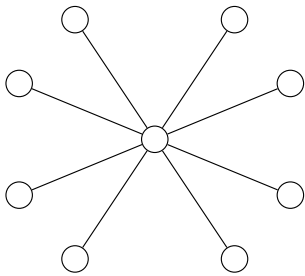
The fact that routing can be used to distribute entanglement means that bounds on entanglement dynamics can also be used to constrain the time taken to perform routing. Previous work has used this idea to bound routing times [9] in terms of entanglement rates [28] and entanglement capacities [29] across system bipartitions. These two major results on entanglement dynamics can be interpreted in terms of bottlenecks, which are of two kinds:

1. Edge bottlenecks: The number of edges across any bipartition of the system constrains the rate at which the bipartitions can be entangled, by the small incremental entangling (SIE) theorem [28]. Thus, in systems with a small number of edges connecting large subsystems, the routing time is large. An example of such system is a d-dimensional array, where the number of edges joining two halves of the system scales as $\Theta(N^{d-1})$, which is asymptotically smaller than the number of vertices on each side ($\Theta(N^d)$). We refer to such a set of edges as an edge bottleneck.

2. Vertex bottlenecks: A vertex bottleneck is a set of vertices, typically much smaller than the rest of the system, that constrains the interactions of the rest of the system. An archetypal example of a vertex bottleneck is found in the star graph S_N, shown in Fig. 1, which is the complete bipartite graph $K_{1,N-1}$ between one (center) vertex and $N-1$ leaf vertices. Intuitively, to implement interactions between qubits on the leaves, we must make use of the central vertex, which forms a bottleneck. On the other hand, this graph does not have a significant edge bottleneck, due to the large number of edges between the central vertex and the leaves. A general vertex bottleneck is depicted in Fig. 2, which represents systems with a tripartite structure. Such systems can be divided into three parts L, C, R such that edges are only present connecting L to C and C to R. The small total entangling property [29] bounds the entanglement capacity, or total amount of entanglement that can be generated across a partition, in terms of the local dimension of the vertex boundary (sites on one side of a partition that are adjacent to sites on the other side), and there-
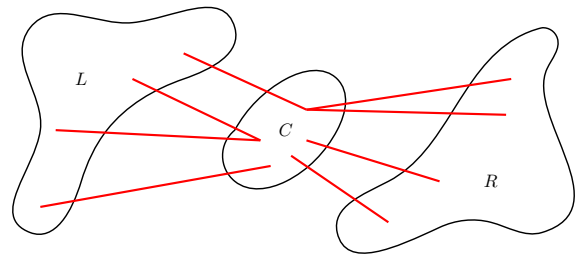
fore constrains the ability to generate entanglement through a vertex bottleneck.

In the gate-based routing model, we make use of 2-qubit gates that act locally, obeying the connectivity constraints of the architecture. A special case of this is classical swap-based routing, where only connectivity-respecting swap gates are permitted. Alternatively, in the more powerful Hamiltonian routing model [9], we are permitted continuous-time evolution by an architecture-respecting Hamiltonian with norm-bounded interactions. In this work, we primarily investigate piecewise time-independent Hamiltonians. The routing time of a model for a given graph G is defined as the time taken to implement a worst-case permutation on G. In the gate-based model, this corresponds to the minimum circuit depth to implement the permutation, while in the Hamiltonian routing model this corresponds to the minimum evolution time required. These routing models are described in more detail in Sec. 2.

Using bounds on the entangling rate [28], tight bounds can be obtained for routing constrained by edge bottlenecks in both the gate-based and continuous-time evolution routing models. However, for routing through vertex bottlenecks, only entanglement bounds in the gate-based model [29] have been found to be tight [9], following the natural assumption that no two gates can interact with the same qubit at the same time. Attempts at a full treatment of the entanglement dynamics in multipartite systems reveal surprisingly more complexity than in the bipartite case. For example, Cubitt *et al.* [31] showed that in a system with a tripartition, entanglement can be generated between the two separated parts without ever entangling the intermediate region with the rest of the system. While this result defies the natural intuition that entanglement must flow between regions of a system, such separable entanglement generation is only possible if the initial state is a special mixed state. For systems starting in a pure state, a notion of entanglement flow between different subsystems was shown by Cubitt *et al.* [30]. However, this result still provides only a trivial lower bound for routing through vertex bottlenecks.

For the star graph, the gate-based quantum routing time is $\Theta(N)$ [9]. However, in the Hamiltonian routing model, which has previously only been shown to be constrained by edge bottlenecks, the best known lower bound on the routing time is $\Omega(1)$ (applying the result of Van Acoleyen *et al.* [28] or a Lieb-Robinson bound [35]). This lower bound intuitively seems loose, since if it were indeed saturable, this would allow for routing in constant time, independent of the system size.

Surprisingly, polynomial speedups for continuous-time evolution over the gate-based model are indeed possible, even in the simple setting of the star graph. For one of our main results, we investigate systems of free (i.e., non-interacting) particles, where the edges of the underlying graph represent two-body hopping terms in the Hamiltonian. In systems of free particles, the gate-based model can be considered to allow architecture-respecting circuits of free-particle gates (i.e., unitaries generated by hopping and on-site interactions). The gate-based routing time on the star graph for free particles is $\Theta(N)$, by the Small Total Entangling lemma [29], applied to circuits of free-particle gates. For these systems, we design a fast Hamiltonian routing protocol that takes time $t = O(\sqrt{N})$ for arbitrary permutations, which demonstrates a $\Omega(\sqrt{N})$ speedup over the gate-based model. We further provide a concise proof that this protocol is optimal (for both free fermions and free bosons), showing a lower bound of $t = \Omega(\sqrt{N})$ on the routing time on the star graph for these systems. We note that this lower bound applies even to time-dependent Hamiltonian routing.

Returning to the more standard setting of qubit systems, we significantly tighten the lower bound on the Hamiltonian routing time for the star graph to $\Omega(N^{1/2-\delta})$ for any $\delta > 0$, when using piecewise time-independent Hamiltonians, or time-dependent Hamiltonians subject to a smoothness condition. This addresses an open question of Bapat *et al.* [9]. The first part of our proof leverages improved commutator bounds on Hamiltonian simulation for Trotter-Suzuki formulas [36] for random input states [37]. This result utilizes the Frobenius norm, allowing us to leverage the fact that vertex-bottlenecked Hamiltonians exhibit tighter commutator scaling in the Frobenius norm compared to the operator norm used in entanglement dynamics bounds [28] or Lieb-Robinson bounds [35]. In contrast to those results, we take advantage of the inherently state-independent task of performing large-scale routing to bound the *average*-case routing time rather than just the worst case, which allows a more precise bound based on the Frobenius norm. The technique behind our proof is inspired by the recent application of "Frobenius light cone" bounds, which demonstrated the hardness of implementing the shift-by-one translation on a ring architecture [38].

This result generalizes to arbitrary vertex bottlenecks, as illustrated in Fig. 2. For an arbitrary tripartition of N_L, N_C, N_R qubits (with $N_L \geq N_R \geq N_C$), we show that the routing time is $\Omega(N_R^{1-\delta}/\sqrt{N_L N_C})$ for any $\delta > 0$.

Our techniques further allow us to prove an upper bound on the entanglement capacity, or the amount of entanglement that can be generated or distributed through a vertex bottleneck. For a tripartition, we define the entanglement capacity for evolution of a state $|\psi\rangle$ by a given Hamiltonian H for time t as the increase in the amount of bipartite entanglement between the L subsystem and the rest of the system. Existing results [28, 30] bound the entanglement capacity as $O(tN_L N_C)$ for any initial state $|\psi\rangle$. For the case where the input state is chosen from a 1-design ensemble, we show an improved upper bound of $O(tN_C\sqrt{N_L}(\sqrt{N_L N_R/N_C})^\delta)$, for any $\delta > 0$, on the entanglement capacity in systems with a tripartition. By averaging over input states, we obtain a tighter bound that depends on Frobenius norms of commutators between terms of the Hamiltonian, rather than the operator norm of the Hamiltonian as used in Refs. [28, 30]. We believe this result could be useful in bounding the time taken for more general quantum information processing tasks beyond routing. For example, our methods may be useful in demonstrating the hardness of implementing specific unitary gates on a quantum processor.

The remainder of the paper is organized as follows. In Sec. 2, we introduce our models and technical background for the routing problem. In Sec. 3.1, we describe a fast routing protocol for systems of free fermions with star-graph connectivity, which performs routing in time $O(\sqrt{N})$. In Sec. 3.2 we discuss the challenges of converting this to a qubit routing protocol. In Sec. 3.3, we show a concise lower bound of $\Omega(\sqrt{N})$ on the routing time for fermions in systems with a vertex bottleneck, including the star graph, which demonstrates that our protocol in Sec. 3.1 is optimal. In Sec. 4, we show our main result bounding the routing time through vertex bottlenecks in systems of qubits. In Sec. 5, we show how similar techniques provide a lower bound on the average entangling capacity in systems with a tripartition. Finally, we discuss implications of our results and some open questions in Sec. 6.

2. PRELIMINARIES

We let $\|\cdot\|$ denote the spectral/operator norm and $\|\cdot\|_p$ the Schatten p-norm. The (normalized) Frobenius norm of an operator O is

$$\|O\|_{\mathrm{F}} := \sqrt{\frac{\mathrm{tr}[O^\dagger O]}{D}}, \tag{1}$$

where D is the total Hilbert space dimension. Note that $\|O\|_{\mathrm{F}} = \|O\|_2/\sqrt{D}$. With this normalization, $\|\mathbf{X}\|_{\mathrm{F}} = 1$ for any Pauli string $\mathbf{X}$.

We specify the connectivity of a quantum system with 2-local interactions in the form of a graph $G = (V, E)$, with the vertices V representing qubits/modes and the edges E representing pairs of qubits/modes between which interactions are allowed. In this paper, we are mostly concerned with graphs with a vertex bottleneck,

as depicted in Fig. 2. Routing is the task of implementing arbitrary permutations from $\mathcal{S}_N$, the group of all permutations of N elements, on qubits labeled from 1 to N. For any $p \in \mathcal{S}_N$, where $p(i)$ represents the destination of the ith qubit, we define the permutation unitary U_p, such that, for any product state $|\psi\rangle = |\psi_1\rangle|\psi_2\rangle \ldots |\psi_N\rangle$,

$$U_p|\psi_1\rangle|\psi_2\rangle \ldots |\psi_N\rangle = |\psi_{p(1)}\rangle|\psi_{p(2)}\rangle \ldots |\psi_{p(N)}\rangle. \quad (2)$$

2.1. Hamiltonian routing

In the Hamiltonian routing model, we consider continuous-time evolution by a Hamiltonian H that respects the connectivity constraints. Such an architecture-respecting evolution has no interaction terms between sites (qubits) that are not connected by an edge in the underlying graph G. Thus, any valid Hamiltonian can be written as

$$H = \sum_{(i,j) \in E} h_{ij} + \sum_{i \in V} h_i, \quad (3)$$

where h_{ij} is only supported on qubits i, j (i.e., acts as identity on all other qubits). We further impose the constraint $\|h_{ij}\| \leq 1$ in order to ensure that the timescale of two-site interactions is comparable to the time required for a 2-qubit gate in the gate-/swap-based routing models. Unitary evolution by a Hamiltonian H is given by the time-evolution operator

$$U(H,t) := \mathcal{T} e^{-\mathrm{i} \int_0^t H(t')\,\mathrm{d}t'}, \quad (4)$$

where $\mathcal{T}$ is the time-ordering operator. When H is time-independent, $U(H,t) = e^{-\mathrm{i}Ht}$.

The *Hamiltonian routing time* is the minimum time to implement a given permutation in this model, defined as

$$\mathrm{hrt}(G,p) := \min_H \{t \text{ s.t. } U(H,t) = U_p\} \quad (5)$$

for any $p \in \mathcal{S}_N$. We define the worst-case Hamiltonian routing time for a graph G as

$$\mathrm{hrt}(G) := \max_{p \in \mathcal{S}_N} \mathrm{hrt}(G,p). \quad (6)$$

In this work, we primarily consider time-independent or piecewise time-independent H.

Definition 2.1 (Piecewise time-independent Hamiltonian). H is said to be piecewise time-independent with minimum segment width Δ if it can be written as

$$H = \sum_{i=0}^{k-1} H_i \cdot \mathbb{I}_{t \in [t_i, t_{i+1})}, \quad (7)$$

where $t_k = t$, H_i is time-independent, $\mathbb{I}_{t \in [t_i, t_{i+1})}$ is an indicator function for $t \in [t_i, t_{i+1})$, and $\forall i, t_{i+1} - t_i \geq \Delta$.

Similarly to the time-dependent case, we define the Hamiltonian routing time for piecewise time-independent Hamiltonians with minimum segment width Δ as

$$\mathrm{hrt}_\Delta(G,p) := \min_H \{t \text{ s.t. } U(H,t) = U_p\} \quad (8)$$

for any $p \in \mathcal{S}_N$. We define the worst-case Hamiltonian routing time for piecewise time-independent Hamiltonians with minimum segment width Δ for a graph G as

$$\mathrm{hrt}_\Delta(G) := \max_{p \in \mathcal{S}_N} \mathrm{hrt}_\Delta(G,p). \quad (9)$$

2.2. Gate-based routing

The gate-based routing model is a variant of the Hamiltonian routing model where we are restricted to performing 1-qubit and 2-qubit gates that respect the connectivity constraints (i.e., no gates are permitted between qubits not connected by an edge). Gates on disjoint pairs of vertices may be applied simultaneously. In this model, the routing time for a given permutation is the minimum depth to implement a given permutation p by an architecture-respecting circuit,

$$\mathrm{rt}(G,p) := \min_U \mathrm{depth}(U) \text{ s.t. } U = U_p. \quad (10)$$

Similar to the Hamiltonian routing model, we also define the *worst-case routing time* for a graph G as

$$\mathrm{rt}(G) := \max_{p \in \mathcal{S}_N} \mathrm{rt}(G,p). \quad (11)$$

The gate-based model generalizes classical routing, which is restricted to circuits composed of architecture-respecting swap gates [4].

2.3. Free-particle routing

An alternative type of Hamiltonian routing is free-particle routing (of bosons or fermions).

In a fermionic model, each vertex of G represents a local fermionic mode, which can be either empty ($|0\rangle$) or occupied by a fermion ($|1\rangle$). Fermionic states and interactions are represented in terms of creation and annihilation operators on each mode ($c_j^\dagger, c_j$) [39].

The annihilation operator acts as follows:

$$c_j|n_0, n_1, \ldots, n_{j-1}, 1, n_{j+1}, \ldots, n_N\rangle$$
$$= (-1)^{\sum_{k=0}^{j-1} n_k} |n_0, n_1, \ldots, 0_j, \ldots, n_N\rangle, \quad (12\mathrm{a})$$
$$c_j|n_0, n_1, \ldots, n_{j-1}, 0, n_{j+1}, \ldots, n_N\rangle = 0. \quad (12\mathrm{b})$$

These operators obey the fermionic anticommutation relations

$$\{c_j, c_k\} = \{c_j^\dagger, c_k^\dagger\} = 0, \quad \{c_j, c_k^\dagger\} = \delta_{j,k}. \quad (13)$$

We also consider routing in systems of free (non-interacting) bosons. Similarly to the fermionic case, states are characterized as Fock states, i.e., in the occupancy number basis $|n_i\rangle$ where n_i is the occupancy of the ith mode (situated on site i). The bosonic creation and annihilation operators are $(a_j^\dagger, a_j)$. The annihilation operator acts as follows:

$$a_j|n_0, n_1, \ldots, n_{j-1}, n_j, n_{j+1}, \ldots, n_N\rangle$$
$$= \sqrt{n_j}|n_0, n_1, \ldots, n_j - 1, \ldots, n_N\rangle. \qquad (14)$$

These operators obey the bosonic commutation relations

$$[a_j, a_k] = [a_j^\dagger, a_k^\dagger] = 0, \ [a_j, a_k^\dagger] = \delta_{j,k}. \qquad (15)$$

Any fermionic or bosonic Hamiltonian can be written as a sum of products of creation and annihilation operators. An architecture-respecting fermionic/bosonic Hamiltonian can be written as a sum of products of these operators on sites that are connected by an edge in G, and have bounded coefficients. In a free (non-interacting) model, the Hamiltonian only contains hopping and on-site terms (in Sec. 6, we comment on the possibility of allowing for pairing). In particular, a free-particle Hamiltonian can be written as

$$H = \sum_{i \in V} h_i c_i^\dagger c_i + \sum_{(i,j) \in E} \left(h_{ij} c_i^\dagger c_j + \text{h.c.} \right), \qquad (16)$$

where $|h_{ij}| \leq 1$, h_i, h_{ij} may have arbitrary time dependence, and c_i are all fermionic or all bosonic.

A permutation $p \in \mathcal{S}_N$ maps each mode on site i to site $p(i)$. Therefore, U_p maps the set of annihilation operators $\{c_1, c_2, \ldots\}$ to $\{c_{p(1)}, c_{p(2)}, \ldots\}$. For particles evolving by a free Hamiltonian with modes b_i, the time-evolved creation and annihilation operators at a given time t can be written as $b_i(t) = \sum_j A_{ij}(t) b_j(0)$ and $b_i^\dagger(t) = \sum_j A_{ij}^*(t) b_j^\dagger(0)$ for some $A_{ij}(t) \in \mathbb{C}$. The dynamics of $A_{ij}(t)$ are the same for both bosons and fermions (and for a single particle). Hence, any protocol that performs routing for free fermions yields an identical routing protocol for free bosons and vice versa.

We define the free-particle routing time as the minimum time to implement a given permutation using an architecture-respecting free-particle Hamiltonian:

$$\text{hrt}_F(G, p) \coloneqq \min_H [t \text{ s.t. } U(H, t) = U_p], \qquad (17)$$

where $U(H, t)$ is the time-evolution operator by H for time t. We define the *worst-case free-particle routing time* for a graph G as

$$\text{hrt}_F(G) \coloneqq \max_{p \in \mathcal{S}_N} \text{rt}_F(G, p). \qquad (18)$$

Since the dynamics of the creation and annihilation operators are identical for bosons and fermions, the routing time does not depend on the particle type.
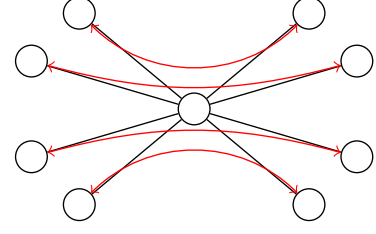

FIG. 3. A global transposition permutation on the star graph. The black lines denote the connectivity constraints. The red lines with arrows indicate pairs of qubits to be swapped.

One can also define a gate-based routing model for free particles. Allowed two-mode gates between mode i and mode j for $(i, j) \in E$ are the unitaries generated by $h_i c_i^\dagger c_i + h_j c_j^\dagger c_j + (h_{ij} c_i^\dagger c_j + \text{h.c.})$ for time-dependent h_i, h_j, and h_{ij}. As in the case of qubits, gates on disjoint pairs of vertices may be applied simultaneously, and application of a single layer of allowed simultaneous gates takes depth (or time) 1. We define the gate-based free-particle routing time $\text{rt}_F(G, p)$ as the minimum depth to implement a given permutation p using an architecture-respecting free-particle circuit (i.e., a circuit composed of free-particle gates). We define the *worst-case* gate-based free-particle routing time for a graph G as $\text{rt}_F(G) \coloneqq \max_{p \in \mathcal{S}_N} \text{rt}_F(G, p)$.

2.4. Transpositions

Definition 2.2 (Graph Tripartition). A tripartition of a graph G is a partitioning of the vertices of G into three sets L, C, R with N_L, N_C, N_R vertices, respectively, such that there are no edges connecting vertices in L to vertices in R. Without loss of generality, L is taken to be the larger of L, R, i.e., $N_L \geq N_R$.

A graph tripartition is depicted in Fig. 2.

Definition 2.3 (Vertex bottleneck). In a graph with a tripartition, C is a vertex bottleneck if $N_C \leq N_R$.

Throughout this paper, we consider tripartitions with a vertex bottleneck, i.e., we assume everywhere that $N_L \geq N_R \geq N_C$. We focus on permutations that perform the maximum possible number N_R of pairs of disjoint swaps between sites through a general vertex bottleneck (Fig. 2). On the star graph S_{N+1}, such a permutation can be obtained by dividing the leaves into two sets L, R of $N/2$ qubits each, and swapping each qubit of L with a corresponding qubit of R. We call this a global transposition, and it is depicted in Fig. 3.

Our investigation of such permutations is motivated by the fact that they can be used at most twice to implement an arbitrary permutation. Any cycle, or set of non-overlapping cycles, can be performed using two global
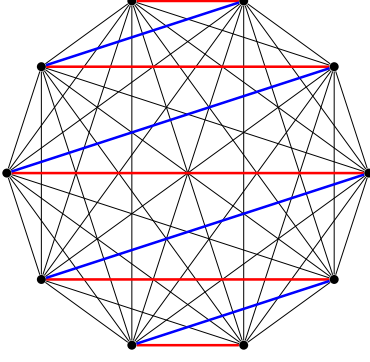
FIG. 4. Any cyclic permutation can be decomposed into two stages of transpositions. For example, a clockwise cycle of all vertices can be achieved by a global transposition between pairs of sites connected by blue edges followed by a global transposition for red edges. To perform an anti-clockwise cycle, simply transpose along the red edges and then the blue edges.

transpositions. Since any permutation can be decomposed into a product of non-overlapping cyclic permutations, arbitrary permutations can be performed using two global transpositions. For example, the cyclic permutation $1 \rightarrow 2, 2 \rightarrow 3, \ldots, 2n \rightarrow 1$, which we denote by $(1\ 2 \ldots 2n)$, can be decomposed as a product of two global transpositions:

$$
\begin{aligned}
(1\ 2 \ldots 2n) &= (2\ 2n)(3\ 2n-1)\cdots(n\ n+2) \\
&\times (1\ 2n)(2\ 2n-1)\cdots(n\ n+1), \quad (19)
\end{aligned}
$$

where the two lines on the right-hand side each consist of non-overlapping pairwise transpositions. This is illustrated in Fig. 4. On the star graph, any permutation can be decomposed into a permutation that acts only on the leaves, followed by a single final swap with the center node. Hence, the time taken for arbitrary permutations on the star graph is predominantly determined by the time to do a global transposition.

2.5. Routing and entanglement distribution

Routing can be used to distribute entanglement, as depicted in Fig. 5. This fact allows us to lower bound the routing time using bounds on the time to generate entanglement. We make use of two such bounds. The small incremental entangling theorem, whose conjecture is attributed to Kitaev in Ref. [32] and which is proved in Ref. [28], bounds the rate at which entanglement can be generated between parts of a system in terms of the local dimension and the operator norm of the Hamiltonian coupling the parts. We restate this result here:

Lemma 2.4 (Small Incremental Entangling (SIE))**.** *Given a Hamiltonian $H = H_{Aa} + H_{Bb} + H_{AB}$ that acts on a system consisting of subsystems a, A, B, b (Fig. 6),*

(a) Initial state with local bell pairs
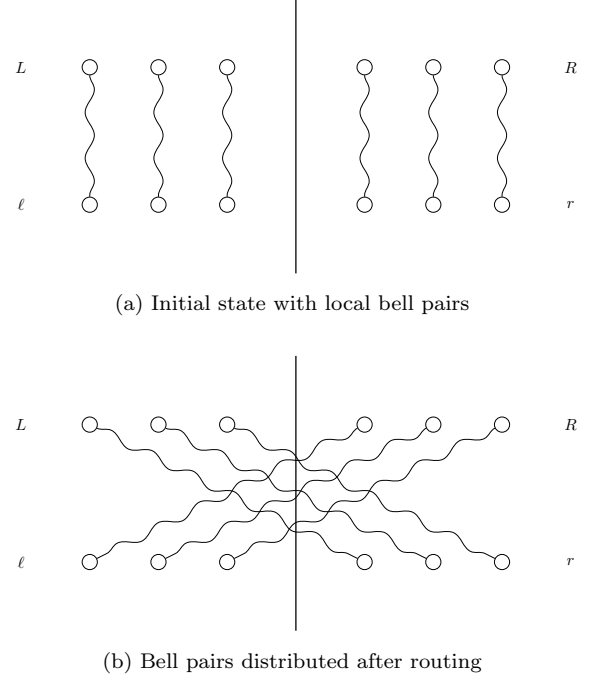

(b) Bell pairs distributed after routing

FIG. 5. Routing enables entanglement distribution. (a) The system has two parts L, R that are initially entangled with hidden auxiliary systems l, r, respectively. (b) On swapping every qubit in L with the corresponding qubit in R, the L system becomes entangled with r, and R becomes entangled with l, so we have generated 6 ebits of entanglement across the cut dividing L and R.
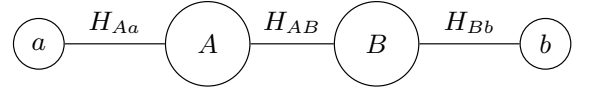

FIG. 6. Partitioning of system for SIE.

for any state ρ,

$$
\frac{\mathrm{d}S_{aA}(\rho)}{\mathrm{d}t} \leq c\|H_{AB}\| \log(d), \quad (20)
$$

where $S_{Aa}(\rho) = -\operatorname{tr}[\rho_A \log \rho_A]$ is the entanglement entropy of ρ across the aA, Bb bipartition, c is some positive constant $(c = 2$ in [40]), and $d = \min\{\dim(A), \dim(B)\}$.

We also make use of the small total entangling property [9, 29], which bounds the total amount of entanglement between parts of a system that can be generated by any unitary evolution, in terms of the local dimension of the parts.

Lemma 2.5 (Small Total Entangling (STE) (Proposition 2 of Ref. [29]))**.** *In a system consisting of subsystems a, A, B, b, for any unitary acting only on A and B, the maximum change in the entanglement entropy of the Aa subsystem is bounded as*

$$
\Delta S_{Aa}(\rho) \leq 2\log(\min\{\dim(A), \dim(B)\}). \quad (21)
$$

In systems of free particles, we quantify bipartite entanglement by the *mode entanglement* of the underlying Fock state [41, 42], which is the von Neumann entropy obtained by tracing out a subset of the modes.

Definition 2.6 (Mode entanglement)**.** For a state of identical particles ρ (i.e., a density matrix in the occupation number basis) with modes partitioned into X and $\bar{X}$, the mode entanglement is

$$S_X(\rho) = -\operatorname{tr}[\rho_X \log \rho_X]. \tag{22}$$

3. ROUTING FREE PARTICLES

3.1. An optimal protocol for routing free particles

Here we consider the problem of routing free particles on a star graph. We give a protocol that can route N free fermions on the star graph in time $\sqrt{N}$. Like the W-state-based routing protocol on the vertex barbell graph (Fig. 9) of Bapat *et al.* [9], this makes use of states spread over multiple sites (here, the leaves) and transfers them one-by-one through the central vertex.

Theorem 3.1. $\operatorname{hrt}_F(S_N) = O(\sqrt{N})$.

Proof. We consider the star graph S_{2N+1}, i.e., a star graph with an even number of leaves. To see that this is sufficient, consider any permutation p on a star graph with an odd number of leaves. We can select some leaf i, and perform a modified permutation p' which is identical to p except when acting on sites i and $p^{-1}(i)$. The permutation p' instead maps $p^{-1}(i)$ to $p(i)$ and leaves i unchanged, and is therefore a permutation on an even number of leaves. By first performing p', and then swapping qubits i and $p(i)$, we can perform the permutation p. Hence it suffices to consider a star graph with an odd number of vertices (i.e., an even number of leaves), since for any permutation p on a star graph with an odd number of leaves $2N + 1$, the routing time is only a constant (3) greater than the routing time for a corresponding permutation on a subgraph with an even number of leaves $2N$.

As explained in Sec. 2, an arbitrary permutation can be implemented by application of at most two global transpositions. For any global transposition, the leaf vertices can be divided into two sets L (left) and R (right) with N leaves on each side, such that the transposition swaps each qubit in L with the corresponding qubit in R. Let the pairs of creation and annihilation operators on each left (right) leaf i be $l_i^\dagger$ and l_i ($r_i^\dagger$ and r_i), respectively. On the center, we denote them as $c^\dagger$ and c.

The left Fourier modes are $f_{l,0}, \ldots, f_{l,N-1}$, where

$$f_{l,k} := \frac{1}{\sqrt{N}} \sum_j e^{-2\pi i k j/N} l_j. \tag{23}$$

Likewise, we denote the right Fourier modes as $f_{r,0}, \ldots, f_{r,N-1}$.

We divide our protocol into N time steps, numbered 0 to $N-1$. At the kth time step, we perform the following operations:

- First, turn on coupling $H_{L,k} = \sum_j e^{-2\pi i j k} c^\dagger l_j + \text{h.c.}$ for time $\frac{\pi}{2\sqrt{N}}$. This Hamiltonian can be equivalently written as $\sqrt{N}(c^\dagger f_{l,k} + \text{h.c.})$ This swaps the center mode with the kth Fourier mode on the left. This is illustrated in Fig. 7a.

- Repeat the same step but with the right leaves using the coupling $H_{R,k} = \sum_j e^{-2\pi i j k} c^\dagger r_j + \text{h.c.}$ This interaction swaps $f_{l,k}$ (which sits at the center following the first step) with $f_{r,k}$ such that $f_{l,k}$ is now on the *right* leaves, $f_{r,k}$ is on the center, and c has been moved to the kth left Fourier mode. This is illustrated in Fig. 7b.

- Finally, perform the coupling $H_{L,k} = \sum_j e^{-2\pi i j k} c^\dagger l_j + \text{h.c.}$ with the left leaves again to swap c and $f_{r,k}$. We have now effectively swapped Fourier modes $f_{l,k}$ and $f_{r,k}$, with c remaining on the center. This is illustrated in Fig. 7a.

Since both Fourier modes and spatial modes form orthonormal bases, swapping all of the corresponding Fourier modes between the left and the right vertices is equivalent to swapping each of the corresponding spatial modes. Since each step of this protocol swapping one Fourier mode at a time takes time $O(\frac{1}{\sqrt{N}})$, the whole protocol takes time $O(\sqrt{N})$. As discussed in Sec. 2, the dynamics of free bosonic and free fermionic creation and annihilation operators are identical. Hence this protocol performs routing for both bosons and fermions. $\square$

3.2. Routing qubits

A natural question is whether a similar protocol exists for qubits. Unfortunately, simply converting the protocol above to a qubit protocol does not work, as we now explain.

First, consider the fermionic version of this protocol. Applying the Jordan-Wigner [43] transform to convert the fermionic operators to qubit operators yields non-local terms, so the generated qubit Hamiltonian does not respect the architecture's geometry and hence does not produce a valid protocol. Even if we permitted ancillary qubits, using an approach such as the Verstraete-Cirac encoding [44] to ensure that we respected locality would introduce a large number of ancillas ($\Theta(N)$) on the central qubit due to the large number of non-commuting terms in the Hamiltonian.

Now we consider whether a bosonic protocol can be converted to a qubit protocol. One approach is to consider the bosonic protocol restricted to an initial state where each mode has occupancy at most 1, and replace the bosonic creation and annihilation operators
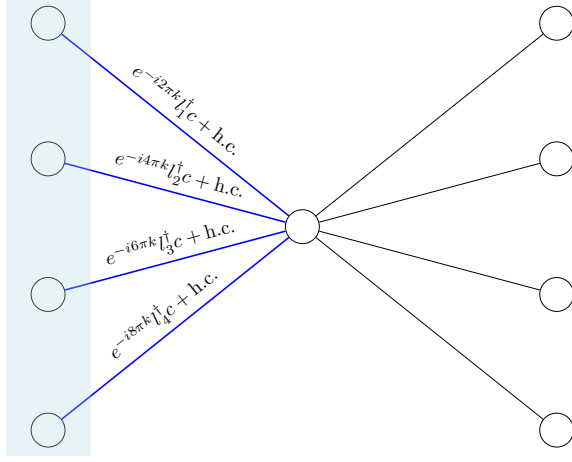
(a) Left swap with center
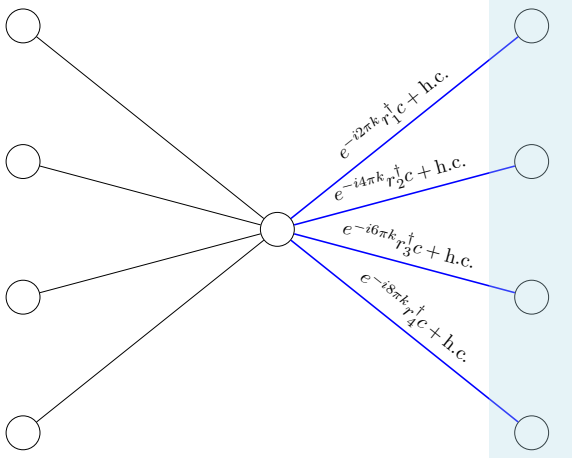
(b) Right swap with center

FIG. 7. Swapping the kth Fourier modes on the left and right sets of leaves.

with spin-1/2 raising and lowering operators, respectively. Another similar strategy to convert a bosonic protocol to a qubit routing protocol is by truncating the bosonic operators to a low (constant) occupancy subspace by replacing the bosonic operators with qudit operators, which may yield an approximate protocol for qubits with ancillas.

Unfortunately, such an approach is infeasible. Consider a bosonic system starting in the uniform superposition of all bitstrings of length N with Hamming weight $N/2$, i.e., the weight $N/2$ Dicke state $|W_{N/2}^N\rangle$. Since every mode has occupancy at most 1, this is a valid initial state for a qubit system as well. As before, we denote the left Fourier modes as $f_{l,0}, \ldots, f_{l,N-1}$, where

$$f_{l,k} := \frac{1}{\sqrt{N}} \sum_j e^{-2\pi i k j/N} l_j. \qquad (24)$$

Likewise, we denote the right Fourier modes as $f_{r,0}, \ldots, f_{r,N-1}$. Then, a simple calculation shows that

the expected occupancy of the zeroth Fourier mode $f_{l,0}$ is $\langle W_{N/2}^N | f_{l,0}^\dagger f_{l,0} | W_{N/2}^N \rangle = \frac{\sqrt{N+2}}{4}$. Hence, if we were to truncate the protocol using qudit operators, it would require qudits of local dimension $\Omega(\sqrt{N})$. Such an approach therefore cannot yield a qubit routing protocol.

Though we have ruled out some obvious approaches, it may still be possible to produce a similar protocol for qubits (or to rule one out entirely), which we believe are interesting directions for future work.

3.3. Lower bound on the free-particle routing time

In this section, we give lower bounds on the time to route N non-interacting particles (fermions or bosons) through a vertex bottleneck.

First, we show a lower bound of $\Omega(N_R/N_C)$ on the gate-based routing model for graphs with a vertex bottleneck. On the star graph, this bound yields $\mathrm{rt}_F(S_N) = \Omega(N)$. This implies that our Hamiltonian-based protocol in Sec. 3.1 obtains a quadratic speedup over gate-based routing.

Then, we show a lower bound on the Hamiltonian routing time $\mathrm{hrt}_F(G)$ for any graph G with a vertex bottleneck. In particular, our proof implies that the free-particle routing time on the star graph is lower-bounded by $\mathrm{rt}_F(S_N) = \Omega(\sqrt{N})$, which implies that the free-particle routing protocol given in Sec. 3.1 is optimal.

To prove these routing-time lower bounds, we first lower bound the time to generate $\Theta(N)$ mode entanglement (Definition 2.6) between two halves of a free-particle system. Analogously to the discussion in Sec. 2.5, the ability of a system to route N non-interacting particles across a partition in time t can be reduced to the problem of generating $\Theta(N)$ mode entanglement in that time—for example, when N modes, each of which is entangled with an ancillary mode as the state $\frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$, are routed across a partition, the amount of entanglement generated across the partition is $\Theta(N)$. As such, a lower bound on the time to generate $\Theta(N)$ entanglement across a vertex bottleneck can be used to lower bound the time to perform routing.

To fix the model more precisely, we assume the system is defined on an architecture graph with a tripartition $V = L \cup C \cup R$ (cf. Fig. 2), where each vertex represents a mode and each edge indicates pairs of modes between which we are allowed to apply a gate. We now show a lower bound on the gate-based routing time for free particles.

Theorem 3.2. *For any graph G with a vertex bottleneck,* $\mathrm{rt}_F(G) = \Omega(N_R/N_C)$.

Proof. We first consider routing in a system of free fermions. As discussed in Sec. 3, due to their identical operator dynamics, any lower bound for routing free fermions also applies to routing free bosons. Consider an

initial state $\rho_{\text{in}} = \frac{1}{2^{N_R}}\mathbb{I}_R \otimes |0\rangle^{N_C+N_L}$. This can be considered as a state where each mode in R is maximally entangled with some reference system that has been traced out. The initial von Neumann entropy of the subsystem R is thus $S_R = N_R$. On routing every mode from R to L, all the entanglement (or entropy) is transferred to L, and $S_R = 0$. Following a similar argument to one presented in Bapat *et al.* [9], by a straightforward application of STE (Lemma 2.5) to free fermions, the maximum change in the von Neumann entropy of R when applying a circuit of depth d is

$$\Delta S_R \leq 2dN_C. \tag{25}$$

On the other hand, to route every mode in R to L requires a total change in the von Neumann entropy of R of $|\Delta S_R| = N_R$. Hence, the minimum circuit depth to perform routing is $\text{rt}_F(G) = \Omega(N_R/N_C)$. $\qquad\square$

Applied to the star graph, this implies a bound of $\text{rt}_F(S_N) = \Omega(N)$, and hence our Hamiltonian routing protocol from Sec. 3.1 obtains a quadratic speedup over gate-based routing.

We now show a lower bound on Hamiltonian free-particle routing in graphs with a vertex bottleneck. As before, each vertex represents a mode and each edge corresponds to a possible two-body "hopping" term in the Hamiltonian. Let the particles evolve under the following non-interacting (time-dependent) Hamiltonian:

$$
\begin{aligned}
H(t) &= H^{(LL)} + H^{(CC)} + H^{(LC)} \\
&= \sum_{i,j\in L} H_{ij}^{(LL)}(t)l_i^\dagger l_j + \sum_{i,j\in C} H_{ij}^{(CC)}(t)c_i^\dagger c_j \\
&\quad + \sum_{\substack{i\in L \\ j\in C}} \left[H_{ij}^{(LC)}(t)l_i^\dagger c_j + \text{h.c.} \right] \\
&\quad + \sum_{\substack{i\in R \\ j\in C}} \left[H_{ij}^{(RC)}(t)r_i^\dagger c_j + \text{h.c.} \right] + \sum_{i,j\in R} H_{ij}^{(RR)}(t)r_i^\dagger r_j,
\end{aligned}
\tag{26}
$$

where the $l_i^\dagger$ (l_i) are the creation (annihilation) operators on L, $c_i^\dagger$ (c_i) are the corresponding operators on C, and $r_i^\dagger$ (r_i) are the corresponding operators on R.

As before, we assume that Hamiltonian terms coupling L and C are bounded by $\|H_{ij}^{(LC)}(t)\| \leq 1$, for all times t.

The following lemma bounds the entangling rate in a free fermionic system. We use this to bound the free fermionic Hamiltonian routing time, and therefore the Hamiltonian routing time for free bosons as well.

Lemma 3.3. *For a free fermionic system defined on a graph with a vertex bottleneck evolving under the free-particle Hamiltonian defined in Eq. (26), the entangling rate between subsystems L and $C \cup R$ is bounded as*

$$\frac{dS_L(\rho(t))}{dt} \leq 4cN_C\sqrt{N_L} \tag{27}$$

Proof. The proof of the theorem follows from applying the Small Incremental Entangling (SIE) theorem (Lemma 2.4) to a free fermionic Hamiltonian acting on a bipartition of the system $(L, C \cup R)$.

By SIE, we are only concerned with $H^{(LC)} = \sum_{\substack{i\in L \\ j\in C}} \left[H_{ij}^{(LC)}(t)l_i^\dagger c_j + \text{h.c.} \right]$. Consider the single normalized mode

$$\tilde{l}_j^\dagger(t) := \frac{1}{\sqrt{\sum_{i\in L}\left|H_{ij}^{(LC)}(t)\right|^2}} \sum_{i\in L} H_{ij}^{(LC)}(t)l_i^\dagger. \tag{28}$$

Then

$$H^{(LC)} = \sum_{j\in C} h_j(t) \tag{29}$$

where $h_j(t) = \sqrt{\sum_{i\in L}\left|H_{ij}^{(LC)}(t)\right|^2}\left[\tilde{l}_j^\dagger(t)c_j + \text{h.c.}\right]$.

Taking the same approach as in the proof of the area law for quasi-adiabatic continuation in [28], we can bound the entangling rate as

$$
\begin{aligned}
\frac{dS_L(\rho(t))}{dt} &= -i\,\text{tr}[H^{(LC)}(t)[\rho(t), \log(\rho_L(t) \otimes \mathbf{1}_C)]] \\
&= -i\sum_{j\in C} \text{tr}[h_j(t)[\rho(t), \log(\rho_L(t) \otimes \mathbf{1}_C)]] \\
&\leq 2c\sum_{j\in C} \|h_j(t)\|,
\end{aligned}
\tag{30}
$$

where $\rho_L(t) = \text{tr}_C[\rho(t)]$, and c is the constant in Lemma 2.4. In the third line, we applied the SIE theorem to bound the contribution of each $h_j(t)$ term separately (cf. Eq. (18) in [28]).

To finish the proof, we bound the norm of $h_j(t)$ as follows:

$$
\begin{aligned}
\|h_j(t)\| &= \left\|\sqrt{\sum_{i\in L}\left|H_{ij}^{(LC)}(t)\right|^2}\left[\tilde{l}_j^\dagger(t)c_j + \text{h.c.}\right]\right\| \\
&\leq 2\sqrt{N_L},
\end{aligned}
\tag{31}
$$

where we have used the fact that, for fermions, $\|\tilde{l}_j^\dagger(t)c_j\| \leq 1$. Plugging this bound on $\|h_j(t)\|$ into Eq. (30) yields

$$\frac{dS_L(\rho(t))}{dt} \leq 4cN_C\sqrt{N_L}, \tag{32}$$

as claimed. $\qquad\square$

Theorem 3.4. *For a system defined on a graph with a vertex bottleneck evolving under the free-particle Hamiltonian defined in Eq. (26),*

$$\text{hrt}_F(G) = \Omega\left(\frac{N_R}{N_C\sqrt{N_L}}\right). \tag{33}$$

Proof. We start with the case of free fermions. By Lemma 3.3, the entangling rate satisfies $\frac{\mathrm{d}S_L(\rho(t))}{\mathrm{d}t} \leq 4cN_C\sqrt{N_L}$. Hence the time to generate $S_L(\rho(t)) = \Theta(N_R)$ ebits of entanglement between subsystems L and $R \cup C$ is lower bounded by

$$t = \Omega\left(\frac{N_R}{\sqrt{N_L}N_C}\right). \quad (34)$$

If every mode in R is initially maximally entangled with an ancillary mode, then routing every mode from R to L requires producing N_R bits of mode entanglement between L and the rest of the system. Hence, the routing time for free fermions (and thus for free bosons as well) is lower bounded as $\Omega\left(\frac{N_R}{\sqrt{N_L}N_C}\right)$ as claimed. $\qquad\square$

We now apply Theorem 3.4 to the case of the star graph. Setting $N_L = N$ and $N_C = 1$ yields a lower bound of $t = \Omega(\sqrt{N})$ on the free-particle routing time, which matches the upper bound given by the free-fermion routing protocol in Sec. 3.1. As such, for free particles, the optimal routing time on the star graph is $\mathrm{hrt}_F(S_N) = \Theta(\sqrt{N})$.

4. LOWER BOUNDS ON QUBIT ROUTING THROUGH VERTEX BOTTLENECKS

We now return to the more common setting of qubits rather than free particles. An architecture-respecting Hamiltonian for a graph with a tripartition can be decomposed as

$$H = H_L + H_C + H_R, \quad (35)$$

where H_L consists of local terms in L, terms coupling different sites in L, and coupling sites in L to sites in C; H_C consists of terms local in C or coupling different sites in C; and H_R consists of terms local in R, coupling different sites in R, and coupling sites in R to sites in C. These terms can be further expanded in terms of Pauli operators $\{X^0, X^1, X^2, X^3\}$:

$$H_L = \sum_{l_1}^{N_L} h_{l_1}^{(L)\alpha_1} X_{l_1}^{\alpha_1} + \sum_{l_1 < l_2}^{N_L, N_L} h_{l_1, l_2}^{(LL)\alpha_1\alpha_2} X_{l_1}^{\alpha_1} X_{l_2}^{\alpha_2}$$
$$+ \sum_{l_1, c_1}^{N_L, N_C} h_{l_1, c_1}^{(LC)\alpha_1\gamma_1} X_{l_1}^{\alpha_1} X_{c_1}^{\gamma_1}, \quad (36a)$$

$$H_C = \sum_{c_1}^{N_C} h_{c_1}^{(C)\gamma_1} X_{c_1}^{\gamma_1} + \sum_{c_1 < c_2}^{N_C, N_C} h_{c_1, c_2}^{(CC)\gamma_1\gamma_2} X_{c_1}^{\gamma_1} X_{c_2}^{\gamma_2}, \quad (36b)$$

$$H_R = \sum_{r_1}^{N_R} h_{r_1}^{(R)\beta_1} X_{r_1}^{\beta_1} + \sum_{r_1 < r_2}^{N_R, N_R} h_{r_1, r_2}^{(RR)\beta_1\beta_2} X_{r_1}^{\beta_1} X_{r_2}^{\beta_2}$$
$$+ \sum_{r_1, c_1}^{N_R, N_C} h_{r_1, c_1}^{(CR)\beta_1\gamma_1} X_{r_1}^{\beta_1} X_{c_1}^{\gamma_1}, \quad (36c)$$

where we use the Einstein summation convention (i.e. summing over repeated indices) for $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2 \in \{1, 2, 3\}$. We assume the coefficients satisfy

$$|h_{l_1}^{(L)\alpha_1}|, |h_{c_1}^{(C)\gamma_1}|, |h_{r_1}^{(R)\beta_1}| \leq \sqrt{N},$$
$$|h_{l_1, l_2}^{(LL)\alpha_1\alpha_2}|, |h_{l_1, c_1}^{(LC)\alpha_1\gamma_1}| \leq 1,$$
$$|h_{c_1, c_2}^{(CC)\gamma_1\gamma_2}|, |h_{r_1, r_2}^{(RR)\beta_1\beta_2}|, |h_{r_1, c_1}^{(CR)\beta_1\gamma_1}| \leq 1 \quad (37)$$

with $N = N_L + N_R + N_C$ and $N_C \leq N_L, N_R$. It is natural to assume even the single-site terms must be bounded by 1. However, some routing models may allow for faster local fields [9]. In this work, we assume a bound of $\sqrt{N}$ for technical reasons to retain our bound on the Frobenius commutator norm in Lemma A.1. Since our graph has a bottleneck, $N_L \geq N_R \geq N_C$. These coefficients may be time-dependent such that H is piecewise time-independent (recall Definition 2.1) with minimum segment width $\propto 1/\sqrt{N_L}$ (this value is due to the limitations of our techniques, as expressed by Eq. (43)). Our Hamiltonian can alternatively be written as $H = H_{LC} + H_R$, where $H_{LC} = H_L + H_C$.

Our main result is a lower bound on the Hamiltonian routing time for any graph with a vertex bottleneck.

Theorem 4.1. *For any graph G with a tripartition into L, C, R with N_L, N_C, N_R vertices, respectively, if C is a vertex bottleneck, then for any constant $\delta \in (0, 1/3]$, if $N_R > 4(2 \times 5^{\frac{1-\delta}{2\delta}})N_C + 2$, there is a constant w_δ such that the routing time for piecewise time-independent Hamiltonians with minimum segment width $\Delta = w_\delta/\sqrt{N_L}$ is*

$$\mathrm{hrt}_\Delta(G) = \Omega\left(\frac{N_R^{1-\delta}}{\sqrt{N_L}N_C}\right). \quad (38)$$

Proof sketch: Our proof consists of two parts. In the first part, in Lemma 4.3, we show that any evolution $U(H, t)$ by an architecture-respecting piecewise time-independent Hamiltonian H with minimum segment width $w_\delta/\sqrt{N_L}$ for time $t \leq c_\delta N_R^{1-\delta}/(\sqrt{N_L}N_C)$ can be well-approximated by an architecture-respecting circuit $\widetilde{U}$ of depth $d \leq N_R/(4N_C) + 2 \times 5^{\frac{1-3\delta}{2\delta}}$ for some constant c_δ. To show this, we first prove Lemma 4.2, which bounds the approximation error of Trotter-Suzuki product-formula circuit approximations to $U(H, t)$ in systems with a tripartition. This lemma is based on the fact that vertex bottlenecks constrain the Frobenius norm of higher-order commutators of terms in any Hamiltonian that respects the tripartition connectivity, which we prove in Lemma A.1. We further make use of recent work [37] that bounds the error of Trotter-Suzuki product-formula circuit approximations in terms of the higher-order-commutator Frobenius norm. In the second part of the proof, we show that such a circuit cannot approximate a desired permutation unitary well. According to STE (Lemma 2.5), the amount of entanglement between L and $C \cup R$ that can be generated by an architecture-respecting circuit of depth d is bounded by $2dN_C$. There

exist permutations p such that the permutation unitary U_p can be used to increase the entanglement between L and R by N_R, an increase which, by STE, requires a circuit of depth $N_R/(2N_C)$. In Lemma 4.4, we show that such a circuit must be far from the target unitary U_p. Now combining the two parts of the proof, since we know that any architecture-respecting evolution $U(H,t)$ for $t \leq c_\delta N_R^{1-\delta}/\sqrt{N_L}N_C$ is well approximated by a circuit $\widetilde{U}$ that cannot perform the permutation p, our main result of a lower bound on the Hamiltonian routing time follows. We defer the formal proof to the end of this section.

In Appendix A, we show the following bound on the error of approximating architecture-respecting evolutions by circuits obtained using the Trotter-Suzuki formula.

Lemma 4.2. *Consider a graph with a vertex bottleneck. Let H be any time-independent Hamiltonian that respects this connectivity. Let $\widetilde{U}$ be the architecture-respecting simulation circuit corresponding to dividing the time-evolution $U = U(H,t)$ into M equal segments, and simulating each by the $(2k)^{\text{th}}$-order Trotter-Suzuki formula. Then, $\widetilde{U}$ has depth $d = 2 \times 5^{k-1}M$ [37] and there exists a function $g(k)$, only dependent on k, such that*

$$\|U - \widetilde{U}\|_\mathrm{F} \leq g(k)\frac{t^{2k+1}}{M^{2k}}\sqrt{N_L^{2k}N_R N_C}. \qquad (39)$$

We use this lemma to show that the circuit obtained from the Trotter-Suzuki formula can be an arbitrarily good approximation of the time-evolution.

Lemma 4.3. *Consider a graph with a vertex bottleneck. Let H be any piecewise time-independent Hamiltonian with minimum segment width Δ that respects the connectivity of a graph with a tripartition. Let $U = U(H,t)$ be the unitary corresponding to evolution by H for time t. Then, for any constant $\delta \in (0, 1/3]$ and $\epsilon > 0$, there exist constants $c_{\delta,\epsilon}, w_{\delta,\epsilon} > 0$ and $k \in \mathbb{N}_+$ such that, for*

$$t \leq \frac{c_{\delta,\epsilon}N_R^{1-\delta}}{\sqrt{N_L}N_C} \qquad (40)$$

and $\Delta > w_{\delta,\epsilon}/\sqrt{N_L}$, there exists an architecture-respecting circuit $\widetilde{U}$ with depth $d \leq N_R/(4N_C) + 2 \times 5^{k-1}$ such that $\|\widetilde{U} - U\|_F \leq \epsilon$.

Proof. We first assume H is time-independent for simplicity, and consider the piecewise case at the end of the proof. Let k be the smallest positive integer such that $\frac{1}{2k+1} \leq \delta$.

Selecting $M = \lceil \frac{1}{4}\left(2 \times 5^{k-1}\right)^{-1}N_R/N_C \rceil$, the circuit approximation $\widetilde{U}$ obtained from Lemma 4.2 has depth $d = 2 \times 5^{k-1}M \leq N_R/(4N_C) + 2 \times 5^{k-1}$. By Lemma 4.2,

$$\|U - \widetilde{U}\|_\mathrm{F}$$
$$\leq 2^{6k}(5^{k-1})^{2k}g(k)$$
$$\times t^{2k+1}\left(\frac{N_C}{N_R}\right)^{2k}\sqrt{N_L^{2k}N_R N_C}. \qquad (41)$$

Bounding the right-hand side by ϵ and solving for t, we obtain

$$t \leq \frac{\left(2^{6k}5^{2k^2-2k}g(k)\epsilon^{-1}\right)^{\frac{-1}{2k+1}}N_R^{1-\frac{1}{2k+1}}}{\sqrt{N_L}N_C}. \qquad (42)$$

This proves the Lemma for time-independent H, because there is a suitable constant $c_{\delta,\epsilon}$ such that Eq. (40) implies Eq. (42), which guarantees $\|U - \widetilde{U}\|_\mathrm{F} \leq \epsilon$.

The above results easily generalize to piecewise time-independent Hamiltonians with minimum segment width $w_{\delta,\epsilon}/\sqrt{N_L}$. The Trotter error formula Eq. (39) does not hold directly because the Hamiltonian changes with time in some of the time windows of duration t/M. Nevertheless,

$$\frac{t}{M} \leq \frac{(8 \times 5^{k-1})c_{\delta,\epsilon}}{N_R^\delta\sqrt{N_L}} < \frac{w_{\delta,\epsilon}}{\sqrt{N_L}} \qquad (43)$$

for constant $w_{\delta,\epsilon}$ determined by $c_{\delta,\epsilon}$, which will be chosen shortly. Since t/M is smaller than the minimum segment width $w_{\delta,\epsilon}/\sqrt{N_L}$, each time window of duration t/M can be split into at most two smaller time windows, each containing time-independent evolution. Applying the Trotter error formula to these possibly smaller time windows, we obtain Eq. (39) again but with an extra prefactor of 2, because there are at most $2M$ new time windows of duration at most t/M. The extra factor of 2 propagates to Eq. (41), so we only need to choose a slightly smaller $c_{\delta,\epsilon}$ than the time-independent case (say, half of its value) to guarantee that Eq. (40) still implies $\|U - \widetilde{U}\|_\mathrm{F} \leq \epsilon$. Hence $w_{\delta,\epsilon} = (4 \times 5^{k-1})c_{\delta,\epsilon}$ suffices. $\square$

We now show that the approximation circuit $\widetilde{U}$ of Lemma 4.3 cannot achieve the permutation that routes the maximum number of qubits through the bottleneck.

Lemma 4.4. *For any graph G with a vertex bottleneck, consider a permutation p that maps m qubits from R to L through the vertex bottleneck and its associated permutation unitary U_p. Then, for any quantum circuit $\widetilde{U}$ with depth $d < m/N_C$,*

$$\|\widetilde{U} - U_p\|_\mathrm{F}^2 > \frac{1}{4}\left(\frac{m - 2dN_C - 1}{N_R}\right)^2. \qquad (44)$$

Proof. First, we bound $\|\widetilde{U} - U_p\|_\mathrm{F}^2$ in terms of the trace distance. For any bit string z of length N_R, we define

$$\rho_z := |z\rangle_R\langle z| \otimes (2^{-N_L-N_C}I)_{LC}. \qquad (45)$$

One can purify the system by a $2^{N_L+N_C}$-dimensional ancilla (labeled by A) so that

$$\rho_z = \mathrm{tr}_A[|\Psi(z)\rangle\langle\Psi(z)|]. \qquad (46)$$

Let $U' = U \otimes I_A$ and $\widetilde{U}' = \widetilde{U} \otimes I_A$. By definition of the Frobenius norm, and introducing $I = 2^{N_L + N_C} \sum_z \rho_z$,

$$\|U - \widetilde{U}\|_F^2$$
$$= \frac{1}{2^{N_R}} \mathrm{tr}\left[\sum_z \rho_z |U - \widetilde{U}|^2\right]$$
$$= \frac{1}{2^{N_R}} \sum_z \langle\Psi(z)| \, |U - \widetilde{U}|^2 \otimes I_A \, |\Psi(z)\rangle$$
$$= \frac{1}{2^{N_R}} \sum_z \left[2 - 2\,\mathrm{Re}(\langle\Psi(z)| U'^\dagger \widetilde{U}' |\Psi(z)\rangle)\right]. \quad (47)$$

Using the fact that $\mathrm{Re}(x) \leq |x|$, and for $x \in [0,1], 1 - x^2 \leq 2(1-x)$, we obtain

$$\sum_z \left[2 - 2\,\mathrm{Re}(\langle\Psi(z)| U'^\dagger \widetilde{U}' |\Psi(z)\rangle)\right]$$
$$\geq \sum_z 2(1 - |\langle\Psi(z)| U'^\dagger \widetilde{U}' |\Psi(z)\rangle|)$$
$$\geq \sum_z 1 - |\langle\Psi(z)| U'^\dagger \widetilde{U}' |\Psi(z)\rangle|^2$$
$$= \frac{1}{4} \sum_z \||U'|\Psi(z)\rangle\langle\Psi(z)|U'^\dagger - \widetilde{U}'|\Psi(z)\rangle\langle\Psi(z)|\widetilde{U}'^\dagger\|_1^2$$
$$(48)$$
$$\geq \frac{1}{4} \sum_z \|\mathrm{tr}_A\left[U'|\Psi(z)\rangle\langle\Psi(z)|U'^\dagger\right.$$
$$\left. - \widetilde{U}'|\Psi(z)\rangle\langle\Psi(z)|\widetilde{U}'^\dagger\right]\|_1^2$$
$$= \frac{1}{4} \sum_z \|\widetilde{U}\rho_z\widetilde{U}^\dagger - U_p\rho_z U_p^\dagger\|_1^2, \quad (49)$$

where in Eq. (48), we have used the relation between fidelity and trace-norm for pure states.

By the small total entangling theorem (Lemma 2.5), $\widetilde{U}$ with depth d can increase the entanglement across the LC, R bipartition by at most $2dN_C$, i.e.,

$$S_R(\widetilde{U}\rho\widetilde{U}^\dagger) - S_R(\rho) \leq 2dN_C, \quad (50)$$

where S_R is the von Neumann entropy of the reduced density matrix on the right N_R qubits, and the vertex boundary of these qubits contains N_C qubits as in Fig. 2.

Observe that ρ_z is pure on R: $S_R(\rho_z) = 0$, so Eq. (50) implies

$$S_R(\widetilde{U}\rho_z\widetilde{U}^\dagger) \leq 2dN_C. \quad (51)$$

On the other hand, U_p routes m identity operators in LC to R, so that

$$S_R(U_p\rho_z U_p^\dagger) = m. \quad (52)$$

Intuitively, two unitaries that produce different changes in the von Neumann entropy of the same initial state

must be distant. By the Fannes-Audenaert Inequality [45] (the second inequality below),

$$\|\widetilde{U}\rho_z\widetilde{U}^\dagger - U_p\rho_z U_p^\dagger\|_1$$
$$\geq \|\mathrm{tr}_{LC}[\widetilde{U}\rho_z\widetilde{U}^\dagger - U_p\rho_z U_p^\dagger]\|_1$$
$$\geq \frac{S_R(U_p\rho_z U_p^\dagger) - S_R(\widetilde{U}\rho_z\widetilde{U}^\dagger) - 1}{\log \dim \mathcal{H}_R}$$
$$\geq \frac{m - 2dN_C - 1}{N_R}. \quad (53)$$

Using this to lower bound Eq. (49) gives the result. $\square$

Combining Lemma 4.3 and Lemma 4.4, we may now prove our main result.

Proof of Theorem 4.1. Let p be a permutation that routes all N_R qubits from R to LC. Let U be the evolution by any architecture-respecting piecewise time-independent Hamiltonian H with minimum segment width $w_{\delta,\epsilon}/\sqrt{N}$ for time $t \leq c_{\delta,\epsilon}N_R^{1-\delta}/\sqrt{N_L}N_C$ for any $\delta > 0$. By Lemma 4.3, U can be ϵ-approximated by an architecture-respecting circuit $\widetilde{U}$ of depth at most $N_R/(4N_C) + 2 \times 5^{k-1}$ such that $\|\widetilde{U} - U\|_F \leq \epsilon$. Now, Lemma 4.4 implies

$$\|\widetilde{U} - U_p\|_F^2 \geq \frac{1}{4}\left(\frac{N_R - 2dN_C - 1}{N_R}\right)^2$$
$$\geq \frac{1}{4}\left(\frac{1}{2} - \frac{2(2 \times 5^{k-1})N_C}{N_R} - \frac{1}{N_R}\right)^2. \quad (54)$$

We thus have

$$\|\widetilde{U} - U_p\|_F \geq \frac{1}{4} - \frac{(2 \times 5^{k-1})N_C}{N_R} - \frac{1}{2N_R}$$
$$\geq \frac{1}{4} - \frac{2(2 \times 5^{k-1})N_C + 1}{2N_R}. \quad (55)$$

From the reverse triangle inequality, we have

$$\|U - U_p\|_F \geq \left|\|\widetilde{U} - U_p\|_F - \|\widetilde{U} - U\|_F\right|$$
$$\geq \frac{1}{4} - \frac{2(2 \times 5^{k-1})N_C + 1}{2N_R} - \epsilon. \quad (56)$$

Since k is the smallest positive integer such that $\frac{1}{2k+1} \leq \delta$,

$$\frac{1}{2(k-1)+1} > \delta$$
$$\implies k - 1 < \frac{1-\delta}{2\delta}. \quad (57)$$

Therefore,

$$N_R > \frac{4(2 \times 5^{\frac{1-\delta}{2\delta}})N_C + 2}{1 - 8\epsilon}$$
$$\implies N_R > \frac{4(2 \times 5^{k-1})N_C + 2}{1 - 8\epsilon}$$
$$\implies \|U - U_p\|_F > \epsilon, \quad (58)$$

and U cannot achieve the desired permutation. Selecting any constant ϵ allows us to pick $w_\delta = w_{\delta,\epsilon}$ for which the theorem holds. $\square$

A straightforward application of Theorem 4.1 to the star graph yields the following result.

Corollary 4.5. *For any $\delta \in (0, 1/3]$, if $N > 8(2 \times 5^{\frac{1-\delta}{2\delta}}) + 4$, the Hamiltonian routing time for the star graph on $N + 1$ vertices with $\Delta = \frac{1}{\sqrt{N}}$ satisfies $\mathrm{hrt}_\Delta(S_{N+1}) = \Omega(N^{\frac{1}{2}-\delta})$.*

5. ENTANGLEMENT FLOW THROUGH BOTTLENECKS

In this section, we investigate the time taken to generate entanglement between subsystems connected by a bottleneck. In a system with a vertex bottleneck, the small total entangling theorem (Lemma 2.5) bounds the amount of entanglement that can be generated between L and RC by a circuit of depth t as $\Delta S_L(\rho) \leq 2tN_C$. On the other hand, the small incremental entangling theorem (Lemma 2.4) [28] and other results on entanglement flow [30] bound the entangling rate in terms of the Hamiltonian operator norm as $\frac{dS_L(\rho)}{dt} \leq c\|H_{LC}\|N_C$. When $N_C \ll N_L$, these two bounds differ greatly. Integrating the bound from Lemma 2.4 for an evolution of time t, we obtain

$$\Delta(S_L(\rho)) \leq cN_L N_C t. \tag{59}$$

Therefore, if this bound were saturable, by evolving for the same amount of time in a Hamiltonian-model rather than a gate-based model, we would obtain a speedup by a factor of N_L in our ability to generate entanglement.

At first, this seems very counter-intuitive. Indeed, in systems of free fermions, we showed a tighter bound of $O(N_C\sqrt{N_L})$ on the entangling rate between L and RC in Lemma 3.3. On the other hand, in systems of free bosons, the entangling rate can be very large, depending on the initial state. Consider a system with just two modes. If the initial state is the Fock state $|N, 0\rangle$, a simple calculation shows that applying a 50/50 beam splitter on the two modes leads to a mode entanglement of $\Theta(\log N)$.

A simple example illustrates that a large increase of the entanglement rate with continuous evolution may be possible for qubits as well. Consider the system depicted in Fig. 8. In a star graph, starting with an initial GHZ state on the leaves and $|0\rangle$ on the center, we can evolve into a GHZ state on the leaves and the center together in time π/N. Since the GHZ state has a von Neumann entropy of 1 between a single qubit and the remaining qubits, this implies that the entangling rate at some time must have been at least linear in N.

Although this system illustrates a worst-case counterexample for the instantaneous entangling rate through a bottleneck, the possibility of the bottleneck restricting our ability to produce entanglement between L and R
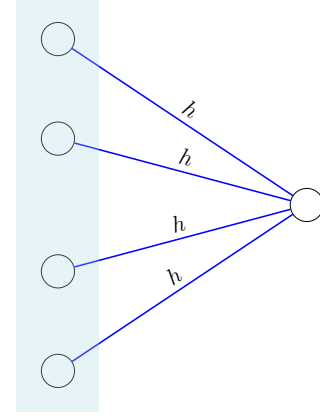

FIG. 8. Linear entangling rate between N leaves of the star graph and the center vertex. If the initial state is a product state of the N-qubit GHZ state $\frac{1}{\sqrt{2}}(|00\ldots0\rangle + |11\ldots1\rangle)$ on the leaves, and $|0\rangle$ on the center, then evolving for time π/N with interaction $h = |1-\rangle\langle1-|_{l_i,c}$ (i.e., the generator of a CNOT) on each edge results in a final $(N+1)$-qubit GHZ state on all the vertices.

remains open. Indeed, if it were always possible to generate fast entanglement, starting in any state, we could use this as a resource to perform routing. The existence of states that allow for fast entangling rates was used in Ref. [9] to obtain a fast routing protocol for the vertex barbell graph, although it also required very fast interactions with ancilla qubits. Our lower bound on the routing time in systems with a vertex bottleneck gives a hint that such states may be rare. Following this hint, we instead look at the entanglement that can be generated *on average* starting with a state drawn from a 1-design.

We define the entanglement capacity $C_\Delta(|\psi\rangle, t) := |S_L(U(H, t)|\psi\rangle\langle\psi|U(H, t)^\dagger - S_L(|\psi\rangle\langle\psi|)|$ as the amount of entanglement between L and RC that can be generated by an architecture-respecting piecewise time-dependent Hamiltonian H of minimum segment width Δ in time t, starting in the state $|\psi\rangle$. Given a pure-state ensemble μ, we define the average entangling capacity as $\mathbb{E}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t)]$. We prove that, in a system with a vertex bottleneck, the average entangling capacity for any 1-design [46–48] scales with N_L as $O(\sqrt{N_L})$. This can be contrasted with the worst-case scaling in N_L of $O(N_L)$ for the entanglement capacity that is obtained by integrating the entangling rate obtained from Lemma 2.4.

Theorem 5.1. *Let μ be a pure-state 1-design and let H be a piecewise time-dependent Hamiltonian with minimum segment width $\Delta = 1/\sqrt{N_L}$ that respects the connectivity constraints of a graph with a vertex bottleneck. Then for any $0 < \delta \leq 1/6$,*

$$\mathbb{E}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t)] = O(1 + tN_C\sqrt{N_L}(N_L N_R/N_C)^\delta). \tag{60}$$

Proof. As in Sec. 4, let $U := U(H, t)$, and $\widetilde{U}$ be the circuit

obtained by the $(2k)^{\text{th}}$ order Trotter-Suzuki formula for U, where k is chosen below. The entangling capacity of H on a particular input pure state $|\psi\rangle$ in time t is

$$
\begin{aligned}
C_\Delta(|\psi\rangle, t) &= |S_L(U|\psi\rangle\langle\psi|U^\dagger) - S_L(|\psi\rangle\langle\psi|)| \\
&= \Big| S_L(U|\psi\rangle\langle\psi|U^\dagger) - S_L(\widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger) \\
&\quad + S_L(\widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger) - S_L(|\psi\rangle\langle\psi|) \Big| \\
&\leq |S_L(U|\psi\rangle\langle\psi|U^\dagger) - S_L(\widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger)| + 2dN_C \\
&\leq 1 + N_L \|U|\psi\rangle\langle\psi|U^\dagger - \widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger\|_1 + 2dN_C,
\end{aligned}
$$
(61)

where in the final line we have applied the Fannes-Audenaert inequality [45].

Note that

$$
\begin{aligned}
&\|U|\psi\rangle\langle\psi|U^\dagger - \widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger\|_1 \\
&= \|(U - \widetilde{U})|\psi\rangle\langle\psi|U^\dagger + \widetilde{U}|\psi\rangle\langle\psi|(U - \widetilde{U})^\dagger\|_1 \\
&\leq \|(U - \widetilde{U})|\psi\rangle\langle\psi|U^\dagger\|_1 + \|\widetilde{U}|\psi\rangle\langle\psi|(U - \widetilde{U})^\dagger\|_1 \quad (62) \\
&= 2\sqrt{\langle\psi|(U - \widetilde{U})^\dagger(U - \widetilde{U})|\psi\rangle},
\end{aligned}
$$
(63)

where in Eq. (62), we have used the triangle inequality, and Eq. (63) follows from the definition of the 1-norm. For states drawn from μ, applying Jensen's inequality and then using the fact that μ is a 1-design,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}\left[2\sqrt{\left\langle\psi\right|(U - \widetilde{U})^\dagger(U - \widetilde{U})\left|\psi\right\rangle}\right] \\
&\leq 2\sqrt{\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}\left[\langle\psi|(U - \widetilde{U})^\dagger(U - \widetilde{U})|\psi\rangle\right]} \\
&= 2\|U - \widetilde{U}\|_{\text{F}}.
\end{aligned}
$$
(64)

Thus, we can bound the expected capacity by the Frobenius distance

$$
\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t)] \leq 1 + 2dN_C + 2N_L\|U - \widetilde{U}\|_{\text{F}}. \quad (65)
$$

As in Lemma 4.3 we first treat the case of time-independent H, and then show how the result generalizes to piecewise time-independent H. Using Lemma 4.2 with

$$
M = \left\lceil t\sqrt{N_L}\,(N_L N_R/N_C)^{\frac{1}{4k+2}}\right\rceil, \quad (66)
$$

we obtain

$$
\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t)] = O(1 + tN_C\sqrt{N_L}\,(N_L N_R/N_C)^{\frac{1}{4k+2}})
$$
(67)

since $d \leq 2 \times 5^{k-1}M$ [37]. The result for the time-independent case follows by choosing k large enough that $1/(4k+2) \leq \delta$.

Since $t/M \leq 1/\sqrt{N_L}$ is smaller than the minimum segment width, the Trotter error formula can be applied, and the result holds in the piecewise time-independent case

with minimum segment width $1/\sqrt{N_L}$. We note that this restriction on the segment width arises for technical reasons (as in Eq. (43)). The number of segments M cannot be chosen to be very large, since this would allow for large d, giving a looser bound on the amount of entanglement generated. Our segment width is thus restricted such that t/M is smaller than the segment width without making M very large.

□

Though there exist examples of states with high entangling rate with a bottleneck (as in Fig. 8), Theorem 5.1 shows that in any 1-design μ there are few such states that can sustain a high entangling rate for long times. In particular, we can show that the entanglement rate for a randomly drawn state over any given period of time is unlikely to be significantly larger than $O(N_C\sqrt{N_L})$. Specifically, applying Markov's inequality and using Theorem 5.1, we obtain

$$
\mathop{\text{Pr}}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t) \geq \Gamma t] = O\left(\frac{1 + tN_C\sqrt{N_L}\,(N_L N_R/N_C)^\delta}{\Gamma t}\right),
$$
(68)

which is vanishingly small when

$$
\Gamma \gg N_C\sqrt{N_L}\left(\frac{N_L N_R}{N_C}\right)^\delta. \quad (69)
$$

For a state $|\psi\rangle$ drawn from μ, let $|\psi(t)\rangle = U(H, t)|\psi\rangle$ be a state along its trajectory under evolution by H. Equation (68) shows that the averaged instantaneous entanglement rate over time,

$$
C_\Delta(|\psi(0)\rangle, T) = \frac{1}{T}\int_0^T \frac{\text{d}S_L(|\psi(t)\rangle\langle\psi(t)|)}{\text{d}t}\text{d}t, \quad (70)
$$

is not greater than $O(N_C\sqrt{N_L}\,(N_L N_R/N_C)^\delta)$ with high probability.

6. DISCUSSION

In this paper, we showed a lower bound that scales as $\Omega(N_R^{1-\delta}/\sqrt{N_L}N_C)$ for any $\delta > 0$ on the routing time in graphs with a vertex bottleneck. For the star graph, this provides a lower bound of $\Omega(\sqrt{N^{1-\delta}})$ on the routing time. We further showed an optimal routing protocol that saturates this bound in systems of free fermions on the star graph.

For graphs with a vertex bottleneck, the best previous lower bound we are aware of is $\Omega(1)$, following from the Small Incremental Theorem [28] or a Lieb-Robinson bound [35]. However, unlike those results, a limitation of our work is our introduction of additional assumptions about the allowed Hamiltonian: piecewise time-independence, absence of ancillas, and bounds on the on-site terms $\|h_i\|$. A straightforward argument can be used to prove a similar lower bound for routing with time-dependent Hamiltonians for which the Frobenius norm

has suitably bounded derivative, by simulating the time-dependent Hamiltonian evolution with a piecewise time-independent evolution, as in [49]. A natural direction for future work is to extend our lower bound to systems with more general Hamiltonians with more general time dependence or unbounded local terms. Our assumption on the piecewise time dependence of the allowed Hamiltonians stems from our use of a Trotterized circuit with Frobenius commutator norm scaling of error in Lemma 4.3. If bounds on the Trotter error for random input states can be extended beyond piecewise time-independent Hamiltonians to time-dependent Hamiltonians whose norm has unbounded derivative, our result could be extended to arbitrary time-dependent Hamiltonian routing as well. Such an approach might also extend our result to Hamiltonians with unbounded local (on-site) terms, since we could move into the interaction picture, removing the on-site terms and adding time-dependence to the non-local terms. We conjecture that a similar lower bound can be proved for more general time-dependent Hamiltonians (whose norms may have large derivatives), and that our assumed bound on the on-site terms can be removed. Another direction for future work is to investigate routing with a piecewise time-independent Hamiltonian of smaller segment width. For technical reasons, our result is restricted to segment width $\propto 1/\sqrt{N_L}$. It is natural to ask whether routing with a piecewise time-independent Hamiltonian of arbitrarily small segment width is equivalent to routing with arbitrary time-dependent Hamiltonians (i.e., whether $\lim_{\Delta \to 0} \mathrm{hrt}_\Delta(G) = \mathrm{hrt}(G)$).

We also assumed that interactions between any two sites in the same partition have bounded norm. This is a natural assumption in systems where interactions between the partitions are physically similar to interactions within the partitions. A possible direction for future work would be to relax this assumption. We expect that this would lead to a looser lower bound on the routing time. Indeed, in systems with ancillas and unbounded interactions within each partition, we can perform routing faster. An example of this is the W-state broadcasting-based protocol of Bapat *et al.* [9], which performs Hamiltonian routing in time $O(\sqrt{N})$ on the vertex barbell graph with $2N + 1$ vertices (Fig. 9), obtaining a speedup over the gate-based model (where routing requires depth $\Theta(N)$). Recent work [50] implies the possibility of even faster routing, in poly($\log N$) time, in the same system. However, speedups obtained in this routing model may not always apply since they require the experimentally challenging capability of fast swap operations between qubits and their ancillas. Our techniques provide an $\Omega(N^{1/2-\delta})$ lower bound in the more realistic setting of the vertex barbell graph without ancillas. Even with bounded 2-site interactions, a related question for future work is to obtain lower bounds that apply when ancillas are permitted.

A more refined analysis might be used to remove the δ dependence of our qubit lower bound. While our proofs
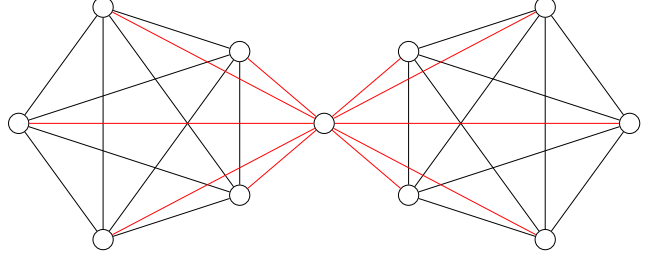

FIG. 9. The vertex barbell graph with $2N + 1$ vertices consists of two copies of the complete graph K_N connected by a common vertex. Here, we depict the vertex barbell with 11 vertices.

hold for arbitrarily small positive constant δ, we conjecture that the result holds with $\delta = 0$. A limitation of our techniques is that our lower bound has a constant factor inversely proportional to δ, so taking the limit as $\delta \to 0$ causes the lower bound to diverge.

We also showed an upper bound on the average entanglement capacity, or average amount of entanglement that can be produced in time t, through a vertex bottleneck. Intuitively, one expects a bottleneck to limit entangling rates, but surprisingly there are states for which fast entanglement with a central qubit is possible. One such example is the W state, which can be used to create constant entanglement with a central vertex in time $O(1/\sqrt{N})$ [9]. The GHZ state allows for an even higher entangling rate, producing a constant amount of entanglement in time $O(1/N)$. Our bound on the average entangling capacity for a 1-design shows that states with an entangling rate higher than $\Omega(\sqrt{N})$ are rare, and constrains their use for tasks such as routing. This result may be interpreted as showing that most states cannot support fast entangling through a vertex bottleneck. An open question is whether this bound can be extended to mixed states as well, which are known to exhibit surprising entanglement dynamics [31]. It would also be interesting to investigate how this bound constrains other quantum information processing tasks based on their requirements of entanglement. Similarly to our routing lower bound, our entanglement capacity upper bound can be extended to time-dependent Hamiltonians whose norm has bounded derivative by approximating their evolution by a piecewise time-dependent one. Natural questions for future work are to extend the entanglement capacity upper bound to Hamiltonians with arbitrary time dependence, with fast on-site interactions and fast interactions within partitions, and with ancillas. We also conjecture that this bound holds with $\delta = 0$.

The existence of states that saturate the worst-case bounds on entangling rate, which depend on the operator norm, necessitates our proof by the average case, using the Frobenius norm. For quantum information tasks involving operators with significant difference in the operator norm and Frobenius norm (as for nested commutators of Hamiltonians with a vertex bottleneck), our re-

sults provide an example where studying the average case yields tighter bounds than the worst case.

Recent work by one of the authors has shown a fast approximate broadcasting (i.e., GHZ encoding) protocol on the complete graph [50]. While this protocol has not been proven to have asymptotically vanishing error, numerics suggest this may be the case. Such a protocol, coupled with fast ancilla interactions, would allow for approximate routing on the vertex barbell in time $\text{poly}(\log N)$. An open question is to resolve whether such broadcasting can be done with vanishing error, or even exactly. The best known exact protocol for routing on the vertex barbell (with fast local ancillas) involves broadcasting into a W state [9], and takes time $O(\sqrt{N})$.

While we developed optimal protocols for routing free particles, a natural open question is whether such protocols can be sped up by allowing interaction terms or even terms that do not preserve particle number, such as $l_i^\dagger c_i^\dagger$. In bosonic systems, introduction of a certain type of interaction term has been shown to provide speedups for the closely related task of state transfer [51].

In Sec. 3.2, we discussed the difficulty of extending our free-particle routing protocols to work for qubits or qudits. A major open question is to obtain fast (sublinear-time) qubit routing protocols through bottlenecks that leverage continuous time evolution. While such results have been obtained on the barbell graph using W states (and we expect that the result of Yin [50] allows for even faster routing using GHZ states), these protocols require the use of ancillas with fast interactions. Currently, we

do not know of any sublinear-time routing protocol for qubits on the star graph.

[1] C. D. Bruzewicz, J. Chiaverini, R. McConnell, and J. M. Sage, Trapped-ion quantum computing: Progress and challenges, Applied Physics Reviews **6**, 021314 (2019), arXiv:1904.04178.

[2] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, Superconducting qubits: Current state of play, Annual Review of Condensed Matter Physics **11**, 369 (2020), arXiv:1905.13641.

[3] A. M. Childs, E. Schoute, and C. M. Unsal, Circuit transformations for quantum architectures, in *14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2019)*, Leibniz International Proceedings in Informatics (LIPIcs), Vol. 135 (2019) pp. 3:1–3:24, arXiv:1902.09102.

[4] N. Alon, F. R. K. Chung, and R. L. Graham, Routing permutations on graphs via matchings, SIAM Journal on Discrete Mathematics **7**, 513–530 (1994).

[5] A. Cowtan, S. Dilkes, R. Duncan, A. Krajenbrink, W. Simmons, and S. Sivarajah, On the qubit routing problem, in *14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2019)*, Leibniz International Proceedings in Informatics (LIPIcs), Vol. 135 (2019) pp. 5:1–5:32, arXiv:1902.08091.

[6] F. Wagner, A. Bärmann, F. Liers, and M. Weissenbäck, Improving quantum computation by optimized qubit routing, Journal of Optimization Theory and Applications **197**, 1161–1194 (2023), arXiv:2206.01294.

[7] G. Li, Y. Ding, and Y. Xie, Tackling the qubit mapping problem for NISQ-era quantum devices, in *Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems* (2019) p. 1001–1014, arXiv:1809.02573.

[8] A. Bapat, A. M. Childs, A. V. Gorshkov, S. King, E. Schoute, and H. Shastri, Quantum routing with fast reversals, Quantum **5**, 533 (2021), arXiv:2103.03264.

[9] A. Bapat, A. M. Childs, A. V. Gorshkov, and E. Schoute, Advantages and limitations of quantum routing, PRX Quantum **4**, 010313 (2023).

[10] D. Devulapalli, E. Schoute, A. Bapat, A. M. Childs, and A. V. Gorshkov, Quantum routing with teleportation, Physical Review Research **6**, 033313 (2024), arXiv:2204.04185.

[11] D. J. Rosenbaum, Optimal quantum circuits for nearest-neighbor architectures, in *8th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2013)*, Leibniz International Proceedings in Informatics (LIPIcs), Vol. 22 (2013) p. 294–307,

arXiv:1205.0036.

[12] F. Hahn, A. Pappa, and J. Eisert, Quantum network routing and local complementation, npj Quantum Information **5**, 1–7 (2019), arXiv:1805.04559.

[13] P. Yuan and S. Zhang, Full characterization of the depth overhead for quantum circuit compilation with arbitrary qubit connectivity constraint (2024), arXiv:2402.02403.

[14] R. Jozsa, Entanglement and quantum computation, in *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose*, edited by S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou, and N. M. J. Woodhouse (Oxford University Press, 1998) pp. 369–379, arXiv:quant-ph/9707034.

[15] R. Jozsa and N. Linden, On the role of entanglement in quantum-computational speed-up, Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **459**, 2011–2032 (2003), arXiv:quant-ph/0201143.

[16] A. Ekert and R. Jozsa, Quantum algorithms: entanglement–enhanced information processing, Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences **356**, 1769–1782 (1998), arXiv:quant-ph/9803072.

[17] D. Bruß and C. Macchiavello, Multipartite entanglement in quantum algorithms, Physical Review A **83**, 052313 (2011), arXiv:1007.4179.

[18] Q. Zhao, Y. Zhou, and A. M. Childs, Entanglement accelerates quantum simulation, to appear in Nature Physics (2025), arXiv:quant-ph/2406.02379.

[19] S. Bravyi, D. Lee, Z. Li, and B. Yoshida, How much entanglement is needed for quantum error correction? (2024), arXiv:2405.01332.

[20] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, Mixed-state entanglement and quantum error correction, Physical Review A **54**, 3824–3851 (1996), arXiv:quant-ph/9604024.

[21] E. Knill and R. Laflamme, Theory of quantum error-correcting codes, Physical Review A **55**, 900–911 (1997), arXiv:quant-ph/9604034.

[22] H. Buhrman, R. Cleve, and W. van Dam, Quantum entanglement and communication complexity, SIAM Journal on Computing **30**, 1829–1841 (2001), arXiv:quant-ph/9705033.

[23] N. Gisin and R. Thew, Quantum communication, Nature Photonics **1**, 165–171 (2007), arXiv:quant-ph/0703255.

[24] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, Quantum entanglement, Reviews of Modern Physics **81**, 865–942 (2009), arxiv:quant-ph/0702225.

[25] C. Degen, F. Reinhard, and P. Cappellaro, Quantum sensing, Reviews of Modern Physics **89**, 035002 (2017), arXiv:1611.02427.

[26] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum-enhanced measurements: Beating the standard quantum limit, Science **306**, 1330–1336 (2004), arXiv:quant-ph/0412078.

[27] J. J. Bollinger, W. M. Itano, D. J. Wineland, and D. J. Heinzen, Optimal frequency measurements with maximally correlated states, Physical Review A **54**, R4649–R4652 (1996).

[28] K. Van Acoleyen, M. Mariën, and F. Verstraete, Entanglement rates and area laws, Physical Review Letters **111**, 170501 (2013), arXiv:1304.5931.

[29] M. Mariën, K. M. R. Audenaert, K. Van Acoleyen, and F. Verstraete, Entanglement rates and the stabil-

[30] ity of the area law for the entanglement entropy, Communications in Mathematical Physics **346**, 35–73 (2016), arXiv:1411.0680.

[30] T. S. Cubitt, F. Verstraete, and J. I. Cirac, Entanglement flow in multipartite systems, Physical Review A **71**, 052308 (2005), arXiv:quant-ph/0404179.

[31] T. S. Cubitt, F. Verstraete, W. Dür, and J. I. Cirac, Separable states can be used to distribute entanglement, Physical Review Letters **91**, 037902 (2003), arXiv:quant-ph/0302168.

[32] S. Bravyi, Upper bounds on entangling rates of bipartite Hamiltonians, Physical Review A **76**, 052319 (2007), arXiv:0704.0964.

[33] W. Dür, G. Vidal, J. I. Cirac, N. Linden, and S. Popescu, Entanglement capabilities of nonlocal Hamiltonians, Physical Review Letters **87**, 137901 (2001), arXiv:quant-ph/0006034.

[34] C. H. Bennett, A. W. Harrow, D. W. Leung, and J. A. Smolin, On the capacities of bipartite Hamiltonians and unitary gates, IEEE Transactions on Information Theory **49**, 1895–1911 (2003), arXiv:quant-ph/0205057.

[35] C.-F. A. Chen, A. Lucas, and C. Yin, Speed limits and locality in many-body quantum dynamics, Reports on Progress in Physics **86**, 116001 (2023), arXiv:2303.07386.

[36] A. M. Childs, Y. Su, M. C. Tran, N. Wiebe, and S. Zhu, Theory of Trotter error with commutator scaling, Physical Review X **11**, 011020 (2021), arXiv:1912.08854.

[37] Q. Zhao, Y. Zhou, A. F. Shaw, T. Li, and A. M. Childs, Hamiltonian simulation with random inputs, Physical Review Letters **129**, 270502 (2022), arXiv:2111.04773.

[38] C. Yin, A. Lucas, and D. T. Stephen, Frobenius light cone and the shift unitary, Physical Review A **109**, 052419 (2024), arXiv:2402.07990.

[39] S. B. Bravyi and A. Y. Kitaev, Fermionic quantum computation, Annals of Physics **298**, 210–226 (2002), arXiv:quant-ph/0003137.

[40] K. M. R. Audenaert, Quantum skew divergence, Journal of Mathematical Physics **55**, 112202 (2014), arXiv:1304.5935.

[41] Y. Shi, Quantum entanglement of identical particles, Physical Review A **67**, 024301 (2003), arXiv:quant-ph/0205069.

[42] P. Zanardi, Quantum entanglement in fermionic lattices, Physical Review A **65**, 042101 (2002), arXiv:quant-ph/0104114.

[43] P. Jordan and E. Wigner, Über das Paulische Äquivalenzverbot, Zeitschrift für Physik **47**, 631–651 (1928).

[44] F. Verstraete and J. I. Cirac, Mapping local Hamiltonians of fermions to local Hamiltonians of spins, Journal of Statistical Mechanics: Theory and Experiment **2005**, P09012 (2005), arXiv:0508353.

[45] K. M. R. Audenaert, A sharp continuity estimate for the von Neumann entropy, Journal of Physics A: Mathematical and Theoretical **40**, 8127 (2007), arXiv:quant-ph/0610146.

[46] A. Ambainis and J. Emerson, Quantum t-designs: t-wise independence in the quantum world, in *22nd Annual IEEE Conference on Computational Complexity* (2007) p. 129–140, arXiv:quant-ph/0701126.

[47] D. Gross, K. Audenaert, and J. Eisert, Evenly distributed unitaries: On the structure of unitary designs, Journal of Mathematical Physics **48**, 052104 (2007), arXiv:quant-

ph/0611002.

[48] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Symmetric informationally complete quantum measurements, Journal of Mathematical Physics **45**, 2171–2180 (2004), arXiv:quant-ph/0310075.

[49] D. W. Berry, A. M. Childs, R. Cleve, R. Kothari, and R. D. Somma, Exponential improvement in precision for simulating sparse Hamiltonians, in *Proceedings of the*

46th Annual ACM Symposium on Theory of Computing (2014) p. 283–292, arXiv:1312.1414.

[50] C. Yin, Fast and accurate greenberger-horne-zeilinger encoding using all-to-all interactions, Physical Review Letters **134**, 130604 (2025), arXiv:2406.10336.

[51] T. V. Vu, T. Kuwahara, and K. Saito, Optimal light cone for macroscopic particle transport in long-range systems: A quantum speed limit approach, Quantum **8**, 1483 (2024), arXiv:2307.01059.

Appendix A: Proof of Lemma 4.2

In this section, we prove Lemma 4.2. First, we upper bound the Frobenius norms of commutators of terms in the Hamiltonian, which allows us to construct low-depth circuit approximations to the Hamiltonian evolution using the Trotter-Suzuki formula.

Lemma A.1. *Given a sequence $\eta_1, \ldots, \eta_{P+Q} \in \{LC, R\}$, let the nested commutator $[H_{\eta_{P+Q}}, \ldots, [H_{\eta_2}, H_{\eta_1}]]$ include H_{LC} a total of P times and H_R a total of Q times. Then there exists an $f(P, Q) > 0$ with no dependence on N_L, N_C, N_R such that*

$$\|[H_{\eta_{P+Q}}, \ldots, [H_{\eta_2}, H_{\eta_1}]]\|_{\mathrm{F}} \leq f(P, Q)\sqrt{N_L^P N_R^Q N_C}. \tag{A1}$$

Proof. The proof strategy is as follows. We first show that the nested commutator can be expressed in terms of sums over Pauli strings with bounded coefficients, using induction to bound the coefficients. The base case is the first-order commutator. We write out terms of the commutator and regroup coefficients of identical Pauli strings that originate from products of different terms, and then bound the resulting coefficient. We take a similar approach to recursively identify the structure of the higher-order nested commutator and bound its coefficients. Finally, we use orthogonality of Pauli strings under the Frobenius norm to compute the bound of Eq. (A1).

For convenience, we re-label the terms in H:

$$H^{(L)} = \sum_{l_1 \in L} h_{l_1}^{(L)\alpha_1} X_{l_1}^{\alpha_1}, \qquad H^{(LL)} = \sum_{\substack{l_1, l_2 \in L \\ l_1 < l_2}} h_{l_1, l_2}^{(LL)\alpha_1 \alpha_2} X_{l_1}^{\alpha_1} X_{l_2}^{\alpha_2}, \qquad H^{(LC)} = \sum_{\substack{l_1 \in L \\ c_1 \in C}} h_{l_1, c_1}^{(LC)\alpha_1 \gamma_1} X_{l_1}^{\alpha_1} X_{c_1}^{\gamma_1},$$

$$H^{(C)} = \sum_{c_1 \in C} h_{c_1}^{(C)\gamma_1} X_{c_1}^{\gamma_1}, \qquad H^{(CC)} = \sum_{\substack{c_1, c_2 \in C \\ c_1 < c_2}} h_{c_1, c_2}^{(CC)\gamma_1 \gamma_2} X_{c_1}^{\gamma_1} X_{c_2}^{\gamma_2},$$

$$H^{(R)} = \sum_{r_1 \in R} h_{r_1}^{(R)\beta_1} X_{r_1}^{\beta_1}, \qquad H^{(RR)} = \sum_{\substack{r_1, r_2 \in R \\ r_1 < r_2}} h_{r_1, r_2}^{(RR)\beta_1 \beta_2} X_{r_1}^{\beta_1} X_{r_2}^{\beta_2}, \qquad H^{(RC)} = \sum_{\substack{r_1 \in R \\ c_1 \in C}} h_{r_1, c_1}^{(CR)\beta_1 \gamma_1} X_{r_1}^{\beta_1} X_{c_1}^{\gamma_1} \tag{A2}$$

(recall that we use the Einstein summation convention for $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2 \in \{1, 2, 3\}$). Here, $\forall i, l_i \in L, r_i \in R, c_i \in C$. We number the vertices in L, C, R uniquely so that these are disjoint sets. We group these terms as

$$H_{LC} = H^{(L)} + H^{(LL)} + H^{(LC)} + H^{(C)} + H^{(CC)}, \tag{A3a}$$

$$H_R = H^{(R)} + H^{(RR)} + H^{(RC)}. \tag{A3b}$$

As mentioned above, we proceed by induction. For any sequence $\eta_1, \ldots, \eta_{P+Q} \in \{LC, R\}$, suppose there are $P \geq 1$ LCs and $Q \geq 1$ Rs (the order does not matter). We will show by induction that the corresponding nested commutator, which we call a (P, Q)-commutator, is of the form

$$[H_{\eta_{P+Q}}, \ldots, [H_{\eta_2}, H_{\eta_1}]] = \sum_{p_1=0}^{P} \sum_{p_2=1}^{P+1-p_1} \sum_{q=0}^{Q} \sum_{l_1 < l_2 < \cdots < l_{p_1}} \sum_{c_1 < c_2 < \cdots < c_{p_2}} \sum_{r_1 < r_2 < \cdots < r_q}$$
$$A_{l_1 \cdots l_{p_1}, r_1 \cdots r_q, c_1 \cdots c_{p_2}; P, Q}^{\alpha_1 \cdots \alpha_{p_1}, \beta_1 \cdots \beta_q, \gamma_1 \cdots \gamma_{p_2}} X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}, \tag{A4}$$

where

$$|A_{l_1 \cdots l_{p_1}, r_1 \cdots r_q, c_1 \cdots c_{p_2}; P, Q}^{\alpha_1 \cdots \alpha_{p_1}, \beta_1 \cdots \beta_q, \gamma_1 \cdots \gamma_{p_2}}| \leq c_{P,Q} N_L^{(P-p_1)/2} N_R^{(Q-q)/2} N_C^{(1-p_2)/2}, \tag{A5}$$

and where $c_{P,Q}$ is a constant that does not depend on N_L, N_C, N_R.

The base case is the first-order commutator

$$[H_{LC}, H_R] = \sum_{c_1, r_1, c_1'} h_{c_1}^{(C)\gamma_1} h_{r_1, c_1'}^{(CR)\beta_1\gamma_1'} X_{r_1}^{\beta_1} [X_{c_1}^{\gamma_1}, X_{c_1'}^{\gamma_1'}] + \sum_{c_1, c_2, r_1, c_1'} h_{c_1, c_2}^{(CC)\gamma_1\gamma_2} h_{r_1, c_1'}^{(CR)\beta_1\gamma_1'} X_{r_1}^{\beta_1} [X_{c_1}^{\gamma_1} X_{c_2}^{\gamma_2}, X_{c_1'}^{\gamma_1'}]$$
$$+ \sum_{l_1, c_1, r_1, c_1'} h_{l_1, c_1}^{(LC)\alpha_1\gamma_1} h_{r_1, c_1'}^{(CR)\beta_1\gamma_1'} X_{r_1}^{\beta_1} X_{l_1}^{\alpha_1} [X_{c_1}^{\gamma_1}, X_{c_1'}^{\gamma_1'}]. \tag{A6}$$

By expanding the commutators and grouping terms based on which sites they are supported (i.e., act non-trivially) on, we can rewrite this in the form

$$[H_{LC}, H_R] = \sum_{r_1, c_1} A_{r_1 c_1}^{\beta_1 \tilde{\gamma}_1} X_{r_1}^{\beta_1} X_{c_1}^{\tilde{\gamma}_1} + \sum_{r_1, c_1, c_2} A_{r_1 c_1 c_2}^{\beta_1 \tilde{\gamma}_1 \tilde{\gamma}_2} X_{r_1}^{\beta_1} X_{c_1}^{\tilde{\gamma}_1} X_{c_2}^{\tilde{\gamma}_2} + \sum_{l_1, r_1, c_1} A_{l_1 r_1 c_1}^{\alpha_1 \beta_1 \tilde{\gamma}_1} X_{l_1}^{\alpha_1} X_{r_1}^{\beta_1} X_{c_1}^{\tilde{\gamma}_1}. \tag{A7}$$

$A_{r_1 c_1}^{\beta_1 \tilde{\gamma}_1}, A_{l_1 r_1 c_1}^{\alpha_1 \beta_1 \tilde{\gamma}_1}, A_{r_1 c_1 c_2}^{\beta_1 \tilde{\gamma}_1 \tilde{\gamma}_2}$ are coefficients of distinct Pauli strings in Eq. (A7). For conciseness, we leave P, Q implicit in our notation for the first-order commutator, where $P = Q = 1$. Also note that, for example, the coefficients $A_{l_1 r_1 c_1}^{\alpha_1 \beta_1 \tilde{\gamma}_1}$ and $A_{r_1 c_1 c_2}^{\beta_1 \tilde{\gamma}_1 \tilde{\gamma}_2}$ are distinguishable since the lower indices come from distinct sets (e.g., $l_1 \in L$ and $r_1 \in R$). Each $A_{r_1 c_1}^{\beta_1 \tilde{\gamma}_1}$ is a sum of coefficients of $X_{r_1}^{\beta} X_{c_1}^{\tilde{\gamma}_1}$ from Eq. (A6):

$$A_{r_1 c_1}^{\beta_1 \tilde{\gamma}_1} = 2 \sum_{\gamma_1, \gamma_1'} h_{c_1}^{(C)\gamma_1} h_{r_1, c_1}^{(CR)\beta_1\gamma_1'} \mathbb{I}[X_{c_1}^{\gamma_1} X_{c_1}^{\gamma_1'} = X_{c_1}^{\tilde{\gamma}_1}], \tag{A8}$$

where $\mathbb{I}[x]$ is the indicator function that returns 0 (1) if x is false (true). For each $A_{r_1 c_1}^{\beta_1 \tilde{\gamma}_1}$, there are four contributing terms from Eq. (A6) (counting the multiplicity factor 2 in Eq. (A8)), obtained by expanding the commutator and making use of the Pauli multiplication rules. The number of such terms is the number of valid indices $(c_1, r_1, c_1', \beta_1, \gamma_1, \gamma_1')$ that, on regrouping Eq. (A6) to the form Eq. (A7), contribute to terms with new indices $(c_1, r_1, \beta_1, \tilde{\gamma}_1)$. A tilde on an index indicates that it is an updated index in the commutator formed by the multiplication of two Paulis as in Eq. (A8), where $\tilde{\gamma}_1$ is the index corresponding to the single-qubit Pauli obtained on multiplying X^{γ_1} and $X^{\gamma_1'}$. Indices without a tilde are unchanged and are identical to those in H_{LC}, H_R.

Likewise, each $A_{l_1 r_1 c_1}^{\alpha_1 \beta_1 \tilde{\gamma}_1}$ is the sum of four coefficients from Eq. (A6), and each $A_{r_1 c_1 c_2}^{\beta_1 \tilde{\gamma}_1 \tilde{\gamma}_2}$ is the sum of eight coefficients. Hence

$$|A_{r_1 c_1}^{\beta_1 \tilde{\gamma}_1}| \le 4 \max |h_{c_1}^{(C)\gamma_1} h_{r_1, c_1}^{(CR)\beta_1\gamma_1}| \le 4\sqrt{N}, \tag{A9a}$$

$$|A_{l_1 r_1 c_1}^{\alpha_1 \beta_1 \tilde{\gamma}_1}| \le 4 \max |h_{l_1, c_1}^{(LC)\alpha_1\gamma_1} h_{r_1, c_1}^{(CR)\beta_1\gamma_1}| \le 4, \tag{A9b}$$

$$|A_{r_1 c_1 c_2}^{\beta_1 \tilde{\gamma}_1 \tilde{\gamma}_2}| \le 8 \max |h_{c_1, c_2}^{(CC)\gamma_1\gamma_2} h_{r_1, c_1'}^{(CR)\beta_1\gamma_1'}| \le 8. \tag{A9c}$$

Now, as different Pauli strings are orthogonal with respect to the Frobenius norm, we have

$$\|[H_{LC}, H_R]\|_{\mathrm{F}}^2 \le \sum_{c_1, r_1} 16N + \sum_{c_1, c_2, r_1} 64 + \sum_{l_1, c_1, r_1} 16$$
$$\le 16N \times N_C \times N_R + 64 N_C^2 \times N_R + 16 N_L \times N_C \times N_R, \tag{A10}$$

satisfying Eq. (A1) for $P = Q = 1$ because we have assumed N_L is larger than N_R, N_C.

We now consider higher-order commutators. We use the induction hypothesis Eq. (A4) to prove that the same equation holds for the $(P+1, Q)$-commutator, i.e., adding one more outermost commutator with H_{LC}. The case $\widetilde{P} = P, \widetilde{Q} = Q + 1$ follows analogously. We use a tilde to denote next-order quantities, e.g., the numbers of LCs and Rs in the new η sequence are

$$\widetilde{P} = P + 1, \qquad\qquad\qquad \widetilde{Q} = Q. \tag{A11}$$

We use $\cdots$ to indicate products over intermediate indices. For example, $X_{l_1}^{\alpha_1} \cdots X_{l_j}^{\alpha_j}$ indicates $\prod_{i=1}^{j} X_{l_i}^{\alpha_i}$. We also use X_j^0 to denote the identity operator for any j. We write $l \notin \overrightarrow{l_{p_1}}$ as shorthand for $l \notin \{l_1, \ldots, l_{p_1}\}$, and similarly for $c \notin \overrightarrow{c_{p_2}}$, etc. We also abbreviate $A_{l_1 \cdots l_{p_1}, r_1 \cdots r_q, c_1 \cdots c_{p_2}; P,Q}^{\alpha_1 \cdots \alpha_{p_1}, \beta_1 \cdots \beta_q, \gamma_1 \cdots \gamma_{p_2}}$ as $A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P,Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}}$. We first compute the commutators of

terms in H_{LC} (whose indices are labeled by primes, e.g., $X_{l_1'}^{\alpha_1'} X_{c_1'}^{\gamma_1'}$) with terms from the (P, Q)-commutator (whose indices are labeled without primes, e.g., $X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}$).

First, we consider the commutator with $H^{(L)}$. Any term $X_{l_1'}^{\alpha_1'}$ in $H^{(L)}$ is non-commuting with $X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}$ iff $\exists j$ such that $l_j = l_1'$ and $\alpha_j \neq \alpha_1'$. Therefore,

$$[H^{(L)}, A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}]$$
$$= \sum_{l_1' \in L} h_{l_1'}^{(L)\alpha_1'} A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} \mathbb{I}[l_1' = l_j] X_{l_1}^{\alpha_1} \cdots X_{l_j}^{0} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}} [X_{l_1'}^{\alpha_1'}, X_{l_j}^{\alpha_j}]. \tag{A12}$$

Now we consider terms in $H^{(LL)}$. For a term $X_{l_1'}^{\alpha_1'} X_{l_2'}^{\alpha_2'}$, we divide the terms from the (P, Q)-commutator which have the form $X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}$ into those with exactly one overlapping site with l_1', l_2' or exactly two overlapping sites. The commutator is thus expressed as

$$[H^{(LL)}, A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}]$$
$$= \sum_{\substack{l_1, l_2 \in L \\ l_1 < l_2}} h_{l_1', l_2'}^{(LL)\alpha_1' \alpha_2'} A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}} \times$$

$$\left(\mathbb{I}[l_1' = l_j] \mathbb{I}[l_2' \notin \overrightarrow{l_{p_1}}] X_{l_1}^{\alpha_1} \cdots X_{l_j}^{0} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{l_2'}^{\alpha_2'} [X_{l_1'}^{\alpha_1'}, X_{l_j}^{\alpha_j}] \right. \tag{A13}$$

$$+ \mathbb{I}[l_1' \notin \overrightarrow{l_{p_1}}] \mathbb{I}[l_2' = l_j] X_{l_1}^{\alpha_1} \cdots X_{l_j}^{0} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{l_1'}^{\alpha_1'} [X_{l_2'}^{\alpha_2'}, X_{l_j}^{\alpha_j}] \tag{A14}$$

$$\left. + \mathbb{I}[l_1' = l_{j_1}] \mathbb{I}[l_2' = l_{j_2}] X_{l_1}^{\alpha_1} \cdots X_{l_{j_1}}^{0} \cdots X_{l_{j_2}}^{0} \cdots X_{l_{p_1}}^{\alpha_{p_1}} [X_{l_1'}^{\alpha_1'} X_{l_2'}^{\alpha_2'}, X_{l_{j_1}}^{\alpha_{j_1}} X_{l_{j_2}}^{\alpha_{j_2}}] \right). \tag{A15}$$

Now we look at the commutator with $H^{(LC)}$. Similarly to $H^{(LL)}$, we divide terms of the order-$(P + Q)$ nested commutator into those that have exactly one or two overlapping sites with each term of $H^{(LC)}$. Therefore,

$$[H^{(LC)}, A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}] = \sum_{\substack{l_1 \in L \\ c_1 \in C}} h_{l_1', c_1'}^{(LC)\alpha_1' \gamma_1'} A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} \times$$

$$\left(\mathbb{I}[l_1' = l_j] \mathbb{I}[c_1' = c_k] X_{l_1}^{\alpha_1} \cdots X_{l_j}^{0} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{c_1}^{\gamma_1} \cdots X_{c_k}^{0} \cdots X_{c_{p_2}}^{\gamma_{p_2}} [X_{l_1'}^{\alpha_1'} X_{c_1'}^{\gamma_1'}, X_{l_j}^{\alpha_j} X_{c_k}^{\gamma_k}] \right. \tag{A16}$$

$$+ \mathbb{I}[l_1' \notin \overrightarrow{l_{p_1}}] \mathbb{I}[c_1' = c_j] X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{l_1'}^{\alpha_1'} X_{c_1}^{\gamma_1} \cdots X_{c_j}^{0} \cdots X_{c_{p_2}}^{\gamma_{p_2}} [X_{c_1'}^{\gamma_1'}, X_{c_j}^{\gamma_j}] \tag{A17}$$

$$\left. + \mathbb{I}[l_1' = l_j] \mathbb{I}[c_1' \notin \overrightarrow{c_{p_2}}] X_{l_1}^{\alpha_1} \cdots X_{l_j}^{0} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}} X_{c_1'}^{\gamma_1'} [X_{l_1'}^{\alpha_1'}, X_{l_j}^{\alpha_j}] \right). \tag{A18}$$

Now we consider the commutator with $H^{(C)}$, which is analogous to the commutator with $H^{(L)}$:

$$[H^{(C)}, A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}]$$
$$= \sum_{c_1' \in C} h_{c_1'}^{(C)\gamma_1'} A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}} \mathbb{I}[c_1' = c_j] X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_j}^{0} \cdots X_{c_{p_2}}^{\gamma_{p_2}} [X_{c_1'}^{\gamma_1'}, X_{c_j}^{\gamma_j}]. \tag{A19}$$

The case for $H^{(CC)}$ is analogous to the commutator with $H^{(LL)}$:

$$[H^{(CC)}, A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} X^{\alpha_1}_{l_1} \cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1} \cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1} \cdots X^{\gamma_{p_2}}_{c_{p_2}}]$$

$$= \sum_{\substack{c_1,c_2 \in C \\ c_1 < c_2}} h^{(CC)\gamma'_1\gamma'_2}_{c'_1,c'_2} A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} X^{\alpha_1}_{l_1} \cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1} \cdots X^{\beta_q}_{r_q} \times$$

$$\left(\mathbb{I}[c'_1 = c_j]\mathbb{I}[c'_2 \notin \overrightarrow{c_{p_2}}] X^{\gamma_1}_{c_1} \cdots X^{0}_{c_j} \cdots X^{\gamma_{p_2}}_{c_{p_2}} X^{\gamma'_2}_{c'_2}[X^{\gamma'_1}_{c'_1}, X^{\gamma_j}_{c_j}] \right. \tag{A20}$$

$$+ \mathbb{I}[c'_1 \notin \overrightarrow{c_{p_2}}]\mathbb{I}[c'_2 = c_j] X^{\gamma_1}_{c_1} \cdots X^{0}_{c_j} \cdots X^{\gamma_{p_2}}_{c_{p_2}} X^{\gamma'_1}_{c'_1}[X^{\gamma'_2}_{c'_2}, X^{\gamma_j}_{c_j}] \tag{A21}$$

$$\left. + \mathbb{I}[c'_1 = c_{j_1}]\mathbb{I}[c'_2 = c_{j_2}] X^{\gamma_1}_{c_1} \cdots X^{0}_{c_{j_1}} \cdots X^{0}_{c_{j_2}} \cdots X^{\gamma_{p_2}}_{c_{p_2}} [X^{\gamma'_1}_{c'_1} X^{\gamma'_2}_{c'_2}, X^{\gamma_{j_1}}_{c_{j_1}} X^{\gamma_{j_2}}_{c_{j_2}}] \right). \tag{A22}$$

Overall, the $(P+1,Q)$-commutator is

$$[H_{LC}, [H_{\eta_{P+Q}}, \cdots, [H_{\eta_2}, H_{\eta_1}]]] = \sum_{p_1=0}^{P} \sum_{p_2=1}^{P+1-p_1} \sum_{q=0}^{Q} \sum_{l_1 < l_2 < \cdots < l_{p_1}} \sum_{c_1 < c_2 < \cdots < c_{p_2}} \sum_{r_1 < r_2 < \cdots < r_q} \left(A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} \right.$$

$$\left. \times [H^{(L)} + H^{(LL)} + H^{(LC)} + H^{(C)} + H^{(CC)}, X^{\alpha_1}_{l_1} \cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1} \cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1} \cdots X^{\gamma_{p_2}}_{c_{p_2}}] \right). \tag{A23}$$

Using Eqs. (A12)–(A22) and grouping terms together, we can rewrite this as

$$[H_{LC}, [H_{\eta_{P+Q}}, \cdots, [H_{\eta_2}, H_{\eta_1}]]] = \sum_{\widetilde{p_1}=0}^{P+1} \sum_{\widetilde{p_2}=1}^{P+2-\widetilde{p_1}} \sum_{q=0}^{Q}$$

$$\sum_{\widetilde{l_1} < \cdots < \widetilde{l_{\widetilde{p_1}}}} \sum_{\widetilde{c_1} < \cdots < \widetilde{c_{\widetilde{p_2}}}} \sum_{r_1 < r_2 < \cdots < r_q} A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$$

$$X^{\widetilde{\alpha_1}}_{\widetilde{l_1}} \cdots X^{\widetilde{\alpha_{\widetilde{p_1}}}}_{\widetilde{l_{\widetilde{p_1}}}} X^{\widetilde{\beta_1}}_{r_1} \cdots X^{\widetilde{\beta_q}}_{r_q} X^{\widetilde{\gamma_1}}_{\widetilde{c_1}} \cdots X^{\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{c_{\widetilde{p_2}}}}. \tag{A24}$$

Each coefficient $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ is a sum of coefficients of the form $A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} \times h$, where h is a coefficient from H_{LC}. We can thus bound $\|A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}\|$ using the induction hypothesis, which bounds $\|A^{\alpha_1\cdots\alpha_{p_1},\beta_1\cdots\beta_q,\gamma_1\cdots\gamma_{p_2}}_{l_1\cdots l_{p_1},r_1\cdots r_q,c_1\cdots c_{p_2};P,Q}\|$, and by counting the number of terms (and their norms) in each of Eqs. (A12)–(A22) that contribute to the $(P+1,Q)$ commutator:

- In Eq. (A12), the number of sites in the support of the commutator is the same as the number of sites in the support of $X^{\alpha_1}_{l_1} \cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1} \cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1} \cdots X^{\gamma_{p_2}}_{c_{p_2}}$, i.e., $\widetilde{p_1} = p_1, \widetilde{p_2} = p_2$. We count the number of terms in the $(P+1,Q)$ commutator that are obtained from Eq. (A12). For a given $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$, we count the number of terms from Eq. (A12) with $l'_1, \overrightarrow{l_{p_1}}$ that would produce a term in the $(P+1,Q)$-commutator with $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$. We see that $\overrightarrow{l_{p_1}}$ must equal $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$, and there are $\widetilde{p_1}$ choices of l'_1 that overlap with a site from $\overrightarrow{l_{p_1}}$. As in Eq. (A9a), we have a multiplicity of 4, obtained by expanding the commutator and from the Pauli multiplication rules. Thus, there are at most $4\widetilde{p_1}$ terms from Eq. (A12) which are grouped together in Eq. (A24) for each coefficient $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$.

- Similar to the analysis of Eq. (A12), we count the number of terms of Eq. (A13) that contribute to a given $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$. In Eq. (A13), the number of sites in L in the support of the commutator increases by 1, so we have $\widetilde{p_1} = p_1 + 1, \widetilde{p_2} = p_2$. There are $\widetilde{p_1}$ choices of l'_2 and $\widetilde{p_1} - 1$ choices of l'_1 that match a given $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$. Hence there are at most $4\widetilde{p_1}(\widetilde{p_1} - 1)$ coefficients from Eq. (A23) that are grouped together in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$.

- Equation (A14) is similar to Eq. (A13). There are at most $4\widetilde{p_1}(\widetilde{p_1}-1)$ coefficients from Eq. (A23) that are grouped together in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$.

- Equation (A15) is non-zero on terms of $H^{(LL)}$ which have two overlapping sites with terms of the (P,Q)-commutator, such that the Paulis commute on one site and anticommute on the other site. An example of this on two sites is $[X_1X_2, X_1Z_2] \propto Y_2$. Once again, we count the number of choices of sites $l_1', l_2', \overrightarrow{l_{p_1}}$ in Eq. (A15) that contribute to a given $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$, i.e., match the sites $\overrightarrow{\widetilde{l_{p_1}}}$. There are at most N_L choices of the site to be removed (l_2'), and $\widetilde{p_1}$ choices of the site to retain (l_1'). For a given Pauli on site l_1', there are 6 commutators that contribute to the coefficient. For example, Y_2 can be obtained from the following commutators: $[X_1X_2, X_1Z_2], [Y_1X_2, Y_1Z_2], [Z_1X_2, Z_1Z_2]$ and the commutators obtained from these by swapping $X_2 \leftrightarrow Z_2$. On expanding each commutator, we obtain two terms. There are thus at most $12N_L\widetilde{p_1}$ contributing terms from Eq. (A15) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1 - 1, \widetilde{p_2} = p_2$.

- Equation (A16) is similar to Eq. (A15). Terms in Eq. (A16) have one overlapping site in C and one overlapping site in L with the (P,Q)-commutator. Similar to Eq. (A15), the number of sites decreases, in either C or L. If the site removed is from C, then there are at most N_C choices of the site to be removed (c_1'), and $\widetilde{p_1}$ choices of the site to retain (l_1'). Likewise, if the site removed is from L, then there are N_L choices of l_1' and $\widetilde{p_2}$ choices of c_1'. There are at most $12N_L\widetilde{p_2}$ terms of Eq. (A16) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1-1, \widetilde{p_2} = p_2$. Likewise, there are at most $12N_C\widetilde{p_1}$ terms of Eq. (A16) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1, \widetilde{p_2} = p_2 - 1$.

- From Eq. (A17), the number of sites in L in the support of the commutator increases by one, while the number of sites in the support from C remains unchanged. Therefore, there are at most $4\widetilde{p_1}\widetilde{p_2}$ terms of Eq. (A17) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1 + 1, \widetilde{p_2} = p_2$.

- Equation (A18) is similar to Eq. (A17), increasing the number of sites from C while leaving the number of sites from L unchanged. There are at most $4\widetilde{p_1}\widetilde{p_2}$ terms of Eq. (A17) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1, \widetilde{p_2} = p_2 + 1$.

- Equation (A19) is similar to Eq. (A12). There are at most $4\widetilde{p_2}$ terms of Eq. (A19) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1, \widetilde{p_2} = p_2$.

- The analysis of Eq. (A20) is similar to that of Eq. (A13). There are at most $4\widetilde{p_2}(\widetilde{p_2}-1)$ terms of Eq. (A13) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1, \widetilde{p_2} = p_2 + 1$.

- The analysis of Eq. (A21) is similar to that of Eq. (A20). There are at most $4\widetilde{p_2}(\widetilde{p_2}-1)$ terms of Eq. (A13) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1, \widetilde{p_2} = p_2 + 1$.

- The analysis of Eq. (A22) is similar to that of Eq. (A15). Equation (A22) is non-zero on terms of $H^{(CC)}$ which have two overlapping sites with terms of the (P,Q)-commutator, such that the Paulis on one site anticommute, and on the other site they commute. We count the number of choices of sites $c_1', c_2', \overrightarrow{c_{p_2}}$ that match with a given $\overrightarrow{\widetilde{c_{p_2}}}$. There are at most N_C choices of the site to remove (c_2') and $\widetilde{p_2}$ choices of the site to retain (c_1'). Hence there are at most $12N_C\widetilde{p_2}$ terms of Eq. (A22) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_2} = p_2 - 1, \widetilde{p_1} = p_1$.

Equation (A24) is the sum of Eqs. (A12)–(A22). Applying the triangle inequality to the cases above, we thus have

$$\|A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}\| \le 4\widetilde{p_1}\max\|A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q}\|\Big|_{p_1=\widetilde{p_1},p_2=\widetilde{p_2}} + \cdots$$

$$\le 4\widetilde{p_1} \times c_{P,Q} N_L{}^{(P-\widetilde{p_1})/2} N_R{}^{(Q-q)/2} N_C{}^{(1-\widetilde{p_2})/2} + \cdots , \tag{A25}$$

where the first term comes from Eq. (A12), and $\cdots$ represents the contributions from the remaining equations. The second line follows from the induction hypothesis [Eq. (A5)]. Adding the terms from the analysis of each of Eqs. (A12)–(A22), we thus have

$$
\begin{aligned}
\|A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}\| \leq & \left[(4\widetilde{p_1}+4\widetilde{p_2}) + \left(8\widetilde{p_1}^2 + 4\widetilde{p_1}\widetilde{p_2} - 8\widetilde{p_1}\right)\sqrt{N_L} \right. \\
& \left. + \left(8\widetilde{p_2}^2 + 4\widetilde{p_1}\widetilde{p_2} - 8\widetilde{p_2}\right)\sqrt{N_C} + (12\widetilde{p_2}+12\widetilde{p_1})\frac{N_C}{\sqrt{N_C}} + (12\widetilde{p_1}+12\widetilde{p_2})\frac{N_L}{\sqrt{N_L}} \right] \\
& \times c_{P,Q} N_L^{(P-\widetilde{p_1})/2} N_R^{(Q-q)/2} N_C^{(1-\widetilde{p_2})/2} \\
\leq & \left[(4\widetilde{p_1}+4\widetilde{p_2}) + \left(8(P+1)^2 + 4P(P+1) - 8\widetilde{p_1}\right)\sqrt{N_L} \right. \\
& \left. + \left(8P^2 + 4P(P+1) - 8\widetilde{p_2}\right)\sqrt{N_C} + (12(P+1)+12P)\frac{N_C}{\sqrt{N_C}} \right. \\
& \left. + (12(P+1)+12P)\frac{N_L}{\sqrt{N_L}} \right] \\
& \times c_{P,Q} N_L^{(P-\widetilde{p_1})/2} N_R^{(Q-q)/2} N_C^{(1-\widetilde{p_2})/2} \\
\leq & (24P^2 + 72P + 32) \times c_{P,Q} N_L^{(P+1-\widetilde{p_1})/2} N_R^{(Q-q)/2} N_C^{(1-\widetilde{p_2})/2} \\
\leq & \; c_{P+1,Q} N_L^{(P+1-\widetilde{p_1})/2} N_R^{(Q-q)/2} N_C^{(1-\widetilde{p_2})/2},
\end{aligned}
\tag{A26}
$$

where, for example, the term $\propto 4\widetilde{p_1}\widetilde{p_2}\sqrt{N_L}$ comes from the contribution of Eq. (A17), with the extra factor $\sqrt{N_L}$ due to $\widetilde{p_1} = p_1 + 1$. The induction hypothesis, Eq. (A5), is thus satisfied for $P+1, Q$ by choosing $c_{P+1,Q} \geq (24P^2 + 72P + 32)c_{P,Q}$.

Now, we use this to show a bound on the commutator norm. Since different Pauli strings are orthogonal, the Frobenius norm is bounded as

$$
\|[H_{\eta_{P+Q}},\cdots,[H_{\eta_2},H_{\eta_1}]]\|_{\mathrm{F}}^2 \leq \sum_{p_1=0}^{P}\sum_{p_2=1}^{P+1-p_1}\sum_{q=0}^{Q}\sum_{l_1<\cdots<l_{p_1}}\sum_{c_1<\cdots<c_{p_2}}\sum_{r_1<r_2<\cdots<r_q}\|A^{\alpha_1\cdots\alpha_{p_1},\beta_1\cdots\beta_q,\gamma_1\cdots\gamma_{p_2}}_{l_1\cdots l_{p_1},c_1\cdots c_{p_2},r_1\cdots r_q;P+1,Q}\|^2
\tag{A27}
$$

$$
\leq \sum_{p_1=0}^{P}\sum_{p_2=1}^{P+1-p_1}\sum_{q=0}^{Q}\sum_{l_1<\cdots<l_{p_1}}\sum_{c_1<\cdots<c_{p_2}}\sum_{r_1<r_2<\cdots<r_q} c_{P+1,Q}^2 N_L^{P-p_1} N_R^{Q-q} N_C^{1-p_2}
\tag{A28}
$$

$$
\leq \sum_{p_1=0}^{P}\sum_{p_2=1}^{P+1-p_1}\sum_{q=0}^{Q}\binom{N_L}{p_1}\binom{N_C}{p_2}\binom{N_R}{q} c_{P+1,Q}^2 N_L^{P-p_1} N_R^{Q-q} N_C^{1-p_2}
\tag{A29}
$$

$$
\leq \sum_{p_1=0}^{P}\sum_{p_2=1}^{P+1-p_1}\sum_{q=0}^{Q}\frac{N_L^{p_1}}{p_1!}\frac{N_C^{p_2}}{p_2!}\frac{N_R^q}{q!} c_{P+1,Q}^2 N_L^{P-p_1} N_R^{Q-q} N_C^{1-p_2}
\tag{A30}
$$

$$
\leq (f(P,Q))^2 N_L^P N_R^Q N_C,
\tag{A31}
$$

where $f(P,Q)$ does not depend on N_L, N_C, N_R. $\qquad\square$

Now we are ready to prove Lemma 4.2, which we restate here.

Lemma A.2. *Consider a graph with a vertex bottleneck. Let H be any time-independent Hamiltonian that respects this connectivity. Let $\widetilde{U}$ be the architecture-respecting simulation circuit corresponding to dividing the time-evolution $U = U(H,t)$ into M equal segments, and simulating each by the $(2k)^{\mathrm{th}}$-order Trotter-Suzuki formula. Then, $\widetilde{U}$ has depth $d = 2 \times 5^{k-1} M$ [37] and there exists a function $g(k)$, only dependent on k, such that*

$$
\|U - \widetilde{U}\|_{\mathrm{F}} \leq g(k)\frac{t^{2k+1}}{M^{2k}}\sqrt{N_L^{2k} N_R N_C}.
\tag{A32}
$$

Proof. We use the decomposition in Eq. (35) to define $\widetilde{U}$ as the simulation circuit corresponding to dividing the time-evolution by H, $U(H,t)$, into M equal segments, and simulating each by the $(2k)$th-order Trotter-Suzuki formula. For example, if $k = 1$,

$$
\widetilde{U} = \left(e^{-\mathrm{i}(t/2M)H_{LC}} e^{-\mathrm{i}(t/M)H_R} e^{-\mathrm{i}(t/2M)H_{LC}} \right)^M.
\tag{A33}
$$

$\widetilde{U}$ consists of $2 \times 5^{k-1}M$ gates each acting on either the left and center qubit, or the right and center qubit [37]. We call $\widetilde{U}$ the circuit approximation. According to Theorem 2 in Ref. [37] (but using the normalized Frobenius norm),

$$\|U - \widetilde{U}\|_{\mathrm{F}} \leq 2(2 \times 5^{k-1})^{2k+1}M \left(\frac{t}{M}\right)^{2k+1} \sum_{\eta_1,\cdots,\eta_{2k+1}\in\{L,R\}} \||[H_{\eta_{2k+1}},\cdots,[H_{\eta_2},H_{\eta_1}]]\||_{\mathrm{F}},$$

$$\leq 2(4 \times 5^{k-1})^{2k+1}M \left(\frac{t}{M}\right)^{2k+1} \max_{\eta_1,\cdots,\eta_{2k+1}\in\{L,R\}} \||[H_{\eta_{2k+1}},\cdots,[H_{\eta_2},H_{\eta_1}]]\||_{\mathrm{F}}. \tag{A34}$$

Let $f^*_{2k+1} \coloneqq \max_{P,Q} f(P,Q)$ s.t. $P + Q = 2k + 1$. Then Lemma A.1 implies the distance is bounded by

$$\|U - \widetilde{U}\|_{\mathrm{F}} \leq 2(4 \times 5^{k-1})^{2k+1}M \left(\frac{t}{M}\right)^{2k+1} \sqrt{f^*_{2k+1}} \sqrt{N_L{}^{2k}N_R N_C}, \tag{A35}$$

which completes the proof with $g(k) = 2(4 \times 5^{k-1})^{2k+1}\sqrt{f^*_{2k+1}}$. $\qquad\square$