

Circuit Folding: Scalable and Graph-Based Circuit Cutting via Modular Structure Exploitation

Shuwen Kan¹, Yanni Li¹, Hao Wang², Sara Mouradian³, Ying Mao¹

¹Fordham University, New York, NY, USA

²Stevens Institute of Technology, Hoboken, NJ, USA

³University of Washington, Seattle, WA, USA

{sk107, yl32, ymao41}@fordham.edu, hwang9@stevens.edu, smouradi@uw.edu

Abstract—Circuit cutting is a promising technique that leverages both quantum and classical computational resources, enabling the practical execution of large quantum circuits on noisy intermediate-scale quantum (NISQ) hardware. Recent approaches typically focus exclusively on either gate cuts or wire cuts, modeling quantum circuits as graphs. However, identifying optimal cutting locations using this representation often results in prohibitively high computational complexity, especially under realistic hardware constraints. In this paper, we introduce CiFOLD, a novel graph-based framework that exploits repetitive modular structures inherent in quantum algorithms, significantly enhancing the scalability and efficiency of circuit cutting. Our approach systematically folds quantum circuits into compact meta-graphs by identifying and merging common gate sequences across entangled qubits, dramatically simplifying subsequent partitioning tasks. We define folding factor and variance to quantify circuit compression and ensure balanced folding. Using these condensed representations, CiFOLD precisely identifies cut locations without exhaustive global graph searches. We perform extensive experiments, comparing CiFOLD with state-of-the-art circuit-cutting techniques. Results demonstrate that CiFOLD achieves superior partition quality and computational efficiency, reducing the number of required cuts by an average of 31.6% and lowering the sampling overhead substantially by 3.55×10^9 . Our findings illustrate that CiFOLD represents a significant advancement toward scalable quantum circuit cutting.

I. INTRODUCTION

Quantum computing holds immense potential by solving complex problems currently intractable for classical computers. However, current noisy intermediate-scale quantum (NISQ) devices suffer from limited qubit counts, connectivity constraints, and gate fidelity issues, severely hindering the practical execution of large-scale quantum circuits.

Circuit cutting has recently emerged as a promising strategy to extend quantum computation beyond current hardware limitations by decomposing large circuits into smaller subcircuits executable on resource-constrained quantum processors. Two main circuit-cutting paradigms: wire cut [1]–[3], which provides exact reconstruction but incurring fixed overhead, and gate cut [4]–[6], which uses quasiprobability decomposition to reconstruct expectation values with reduced sampling overhead compared to wire cut.

Despite the complementary nature of gate- and wire-cutting methods, their integration into a unified framework remains largely unexplored. Recent studies have begun examining the advantages of combining these two approaches [7], [8], but finding optimal cut points within a hybrid gate-wire partition-

ing is inherently challenging. Existing graph-based techniques for wire cutting represent each two-qubit gate as a node [2], [9], whereas gate cutting adopts a simpler representation, using each qubit as a node [4]. When integrating both gate and wire cuts into one framework, the resulting graph structure effectively requires twice as many nodes compared to the wire-cut-only scenario. Consequently, solver-based methods designed to identify optimal partitions (e.g., Satisfiability Modulo Theories(SMT) solvers used in [7] or Integer Linear Programming(ILP) solvers employed in [8]) face significant runtime overhead. For instance, Brandhofer et al. [7] imposed a one-hour SMT solver timeout for circuits of up to 40 qubits, while Pawar et al. [8], using an improved ILP model, reported solution times of up to 1800 seconds for a 20-qubit quantum Fourier transform (QFT) circuit. They suffer from significant scalability issues and are impractical in reality.

Motivated by the fact that many quantum algorithms follow fixed structural patterns to solve well-defined problems such as Grover’s algorithm [10], quantum phase estimation [11], quantum arithmetic operation [12], Bernstein-Vazirani algorithm [13] and Shor’s algorithm [14]. As problem size increases, these algorithms typically scale by adding more qubits while retaining the same circuit structure. While this modularity offers an opportunity for optimization, current approaches often treat each gate as unique, leading to a combinatorial explosion in cut-point searches that is both non-optimal and computationally expensive. This results in non-optimal or time-consuming methods that fail to scale efficiently. To address these limitations, this work proposes CiFOLD, a novel folding-based circuit-cutting framework designed to harness the inherent modularity of many quantum algorithms. Rather than searching the entire circuit for cutting points, CiFOLD identifies recurring gate sequences and exploits these structural redundancies to minimize the number of required cuts. By incorporating self-adaptive algorithms tailored to detect, from qubit-level, in parallel, and leverage repeated modules, CiFOLD significantly reduces the exponential overhead commonly associated with circuit partitioning. Experimental results on standard quantum benchmarks demonstrate how our approach not only lowers the computational cost but also preserves accuracy, offering a practical path forward for large-scale quantum computations. This work introduces several key innovations:

- **Graph-Based Circuit Folding:** We present a novel graph-based framework for circuit partitioning that models quantum circuits as directed graphs. By applying folding techniques to identify recurring gate sets, CiFOLD constructs a meta-graph that reduces circuit complexity and improves partitioning efficiency.
- **Folding Metrics:** We define the folding factor and folding variance as quantitative measures of circuit folding efficiency, providing a systematic evaluation of the gains achieved through structured circuit decomposition.
- **Adaptive Partitioning:** CiFOLD leverages the meta-graph structure to guide the search for optimal partition points, efficiently reducing the number of subcircuit executions while maintaining fidelity. This adaptive approach ensures flexibility across varying hardware constraints and qubit resources.
- **Implementation and Evaluation:** We implement end-to-end circuit cutting pipeline using CiFOLD and evaluate it across a diverse set of quantum circuit benchmarks. CiFOLD improves relative fidelity, ranging from 5.3% to 61.2% across seven IBM backend emulators, achieves runtimes below 1-second even for complex circuits, and reduces required cuts by an average of 31.6%, leading to a sampling overhead reduction of 3.55×10^9 .

II. RELATED WORK AND BACKGROUND

Quantum circuit cutting broadly falls into two categories: theoretical explorations, primarily investigating bipartite cuts [3], [15]–[22], and practical, implementation-driven studies aimed at efficient real-world deployment [1], [2], [4], [9], [23]–[26]. Early theoretical work employed quasiprobability decomposition, analyzing overhead scaling and techniques such as parallel cuts and classical communication. Practical implementations like CutQC [2], TensorQC [1], and FitCut [9] introduced automated pipelines addressing classical computation bottlenecks and hardware-aware optimization.

In contrast to prior work, CiFOLD framework uniquely exploits qubit-level structural repetition within quantum circuits, using a graph-based folding method to generate compact meta-graphs. This approach reduces computational overhead, enables parallel subcircuit processing, and efficiently optimizes cut selection for both quantum and classical resources.

A. Circuit Cutting

Quantum circuit cutting allows large quantum circuits to be divided into smaller subcircuits that can be executed independently on quantum hardware and later reconstructed using classical postprocessing. Two main techniques have been developed for this purpose: gate cutting and wire cutting, as shown in Figure 1.

Gate cut: employs quasiprobability simulation to reconstruct the expectation value of the original circuit. It replaces a two-qubit gate with local single-qubit operations, denoted by s_i as shown in Fig 1, along with classical postprocessing [15]. Each gate cut results in separately executable subcircuits,

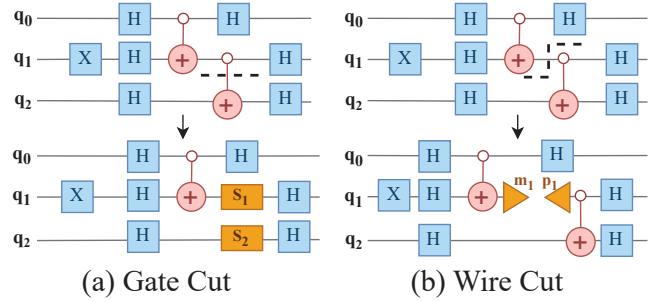


Fig. 1: Illustration of Gate Cut and Wire Cut

which are individually sampled according to their corresponding quasiprobability distribution. **Wire cuts:** A wire cut severs a quantum wire by inserting measurement and preparation operations that effectively separate a circuit into independently executable fragments. As shown in Fig. 1(b), when two wires are cut, m_1 and m_2 correspond to measurement channels in the $\{I, X, Y, Z\}$ bases, while p_1 and p_2 represent state preparations in the $\{|0\rangle, |1\rangle, |i\rangle, |+\rangle\}$ states for a distinct computation fragment [2]. Consequently, one cut entails executing three different upstream subcircuits (since I and Z are identical) and four different downstream subcircuits. The final outputs are combined via 16 Kronecker products of the respective probability distributions to achieve exact reconstruction.

B. End-to-End Circuit Cutting pipeline

Circuit cutting is a hybrid quantum-classical technique that redistributes computational workload between quantum hardware and classical postprocessing. To enable its practical deployment, an end-to-end workflow is required that coordinates quantum execution with classical preprocessing and reconstruction. This workflow consists of three key stages:

1) Cut Point Identification: The first step is to determine the cut locations within the quantum circuit by jointly considering hardware constraints and the requirements of circuit cutting protocols. While the maximum number of qubits per subcircuit is typically the primary constraint, other hardware limitations—such as gate depth restrictions imposed by coherence time and gate fidelity—must also be included in the cut selection process.

The cost model must reflect the specific circuit cutting protocol in use. For example, gate cuts and wire cuts differ significantly in their reconstruction overhead and operational requirements. Techniques like parallel cutting can reduce joint sampling overhead between subcircuits [6], [17], but they introduce additional constraints and trade-offs that must be explicitly accounted for in the model. The overall objective is to minimize the total sampling overhead required for reconstruction, which can scale exponentially with the number and type of cuts. Therefore, identifying an optimal cut strategy requires careful alignment between hardware feasibility and protocol-specific cost modeling.

2) Subcircuit Execution: The second step involves compiling and executing subcircuits. Wire cutting uses a fixed number of shots per subcircuit, while gate cutting requires

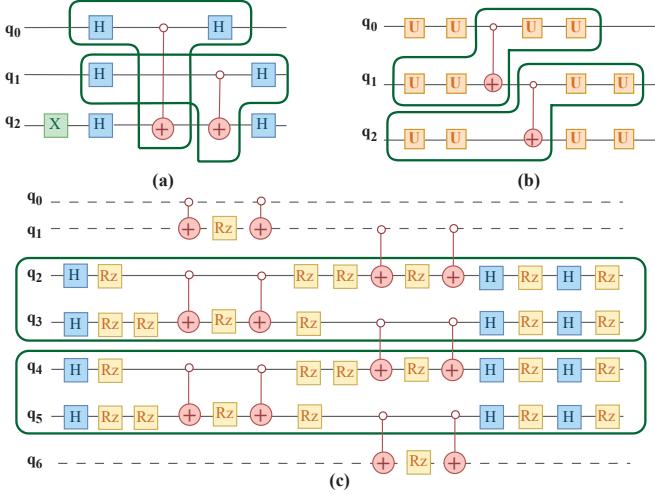


Fig. 2: Circuits with repeated patterns: (a) Bernstein Vazirani (BV); (b) Hardware efficient ansatz (HWEA); (c) ISING

uneven shot allocation based on quasiprobability weights to ensure statistical accuracy.

3) Classical Postprocessing: Following quantum execution, classical reconstruction is performed to assemble either the full probability distribution (in the case of wire cutting) or the expectation value of specified observables (in the case of gate cutting). This reconstruction step follows the protocol determined during the initial cut planning phase.

III. CHALLENGES AND OPPORTUNITIES

Insight 1: Cut Finding is Non Trivial — Despite its critical role, the problem of cut point finding is often overlooked or assumed to be predefined in many circuit cutting studies [3], [5], [6], [27]. Theoretical works frequently operate under the assumption that optimal cuts are known a priori. In practice, determining where to cut is a complex optimization task that cannot be easily addressed using well established graph algorithms. To address this, existing implementations typically formulate cut finding as a constrained optimization problem, employing solvers such as Mixed Integer Programming (MIP), Boolean Satisfiability (SAT), or Integer Linear Programming (ILP) to model sampling overhead and search for globally optimal partitions [2], [7], [8].

The complexity of the graph model differs greatly between gate and wire cutting. Gate cutting’s graph representations map nodes to qubits, enabling heuristic-based merging strategies [4]. In contrast, wire cuts model nodes as two-qubit gates, making valid partitioning under qubit constraints a factorial-complexity problem [2], [9]. Hybrid gate-wire modeling further doubles graph size and increases solver complexity. Even with optimized ILP formulations [8], solving times on practical circuits can exceed 1800 seconds, underscoring the need for more scalable and hardware-aware heuristics.

These challenges present a key opportunity for future research: developing efficient, approximate cut-finding algorithms that balance accuracy with runtime, while still accounting for hardware constraints and sampling overhead.

Insight 2: Exploiting Quantum Circuit Modularity — Quantum algorithms are typically implemented through quantum circuits whose size scales with the complexity of the target problem. These circuits often exhibit recurring patterns across qubit registers or layers. Benchmark suites QASMBench [28] and SupermarQ [29] show how numerous algorithm families (e.g., quantum arithmetic [12], simulation [30], machine learning [31], hidden subgroup [13], search [10], optimization [32] and variational methods [33]) reuse the same structural blocks as qubit counts increase. Figure 2 shows three examples with repeated modules (in green).

Ideally, leveraging modular patterns keeps cutting a 100-qubit BV circuit as manageable as a 10-qubit one. In contrast, general-purpose frameworks that treat each gate independently scale poorly with circuit size. By identifying and exploiting this modularity, e.g., structural repetitions, our approach targets only the unique substructures within the circuit, effectively amortizing the cut-finding cost. This reduces the need for exhaustive search across the entire circuit graph, mitigating the factorial growth in complexity. Instead of treating each gate instance independently, our system constructs a compressed circuit information graph and reuses cutting decisions across all instances of the same module.

Insight 3: Limitations of Classical Graph Algorithms — Classical graph algorithms, such as subgraph-isomorphism solvers or frequent subgraph mining (FSM) techniques, are not well-suited for quantum circuit cutting. First, subgraph-isomorphism approaches require predefined patterns, which are not available in the circuit cutting context. Additionally, conventional graph properties, like node count or topology, do not directly map to quantum hardware constraints or sampling overhead. Second, FSM algorithms typically identify recurring subgraphs within datasets and thus require significant modifications to handle single large graphs, as in circuit cutting. Even with adaptations, FSM methods such as gSpan [34] or MoFa [35] exhibit exponential complexity due to the combinatorial explosion of candidate subgraphs, making them impractical for circuits beyond 100–200 gates.

Consequently, these classical methods fail to effectively align with the practical constraints and cost models necessary for scalable and efficient quantum circuit partitioning.

IV. CiFOLD SOLUTION DESIGN

To address the challenges and limitations, we propose CiFOLD, a scalable and graph-based circuit cutting and folding framework. In this section, we formulate circuit cutting as a constrained graph partitioning problem, introducing a novel weighted graph representation that unifies gate and wire cuts while quantifying sampling overhead as edge weights. CiFOLD hinges on two pillars: (1) circuit folding, a systematic compression technique that identifies recurrent gate sequences across entangled qubits to construct a compact meta-graph; and (2) adaptive partitioning, which leverages the meta-graph’s hierarchical structure to guide minimal-overhead cut identification without exhaustive searches. Central to our approach are the folding factor and folding variance metrics, which

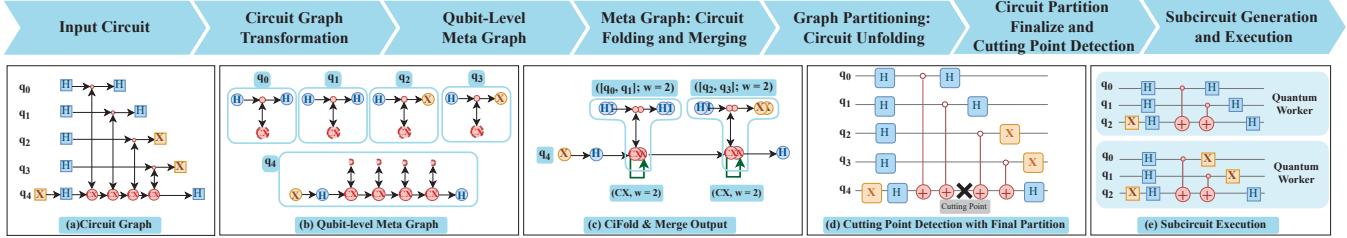


Fig. 3: CiFOLD Framework Overview

rigorously assess compression efficiency and structural balance, enabling automated optimization of the folding process. From a system design perspective, we detail a parallelized workflow that combines dynamic program analysis for pattern detection (via Longest Consecutive Common Subsequence) with Weisfeiler-Lehman graph hashing for fast structural equivalence checks, ensuring scalability to large-scale circuits.

A. Problem Formulation

We represent a quantum circuit as a directed weighted graph $G = (V, E, w)$, where:

- V is the set of nodes representing qubit operands.
- E is the set of edges, each $(u, v) \in E$ corresponding to a two-qubit gate or a wire (sequential gate on same qubit). The directed edges enforce the sequential gate ordering on each qubit, and bidirectional edges represent two-qubit interactions.
- $w : E \rightarrow \mathbb{R}^+$ assigns sampling overhead as edge weights.

Cutting Overhead: Wire cuts have a fixed weight of 16. For gate cuts, the overhead depends on the entangling strength. A controlled-CZ rotation with angle θ incurs: $w(e) = (1 + 2 \sin(\frac{\theta}{2}))^2$. Typical controlled gates (X, Y, Z) have overhead 9, while SWAP has 49.

Partitioning: A valid partitioning $\mathcal{P} = \{P_1, \dots, P_k\}$ satisfies: $\bigcup_{i=1}^k P_i = V$, $P_i \cap P_j = \emptyset \quad \forall i \neq j$. Each partition contains nodes $V_i \subseteq V$ and edges $E_i = \{(u, v) \in E \mid u, v \in V_i\}$. The cut set is: $E_{\text{cut}} = \{(u, v) \in E \mid u \in P_i, v \in P_j, i \neq j\}$. The total sampling overhead is:

$$T_{\text{total}} = \prod_{(u, v) \in E_{\text{cut}}} w(u, v). \quad (1)$$

B. Meta-Graph and Folding Metrics

Given repeated substructures, we construct a meta-graph $G' = (V', E')$ where each $v' \in V'$ and $e' \in E'$ aggregates multiple instances from G . Each meta-node/edge is assigned a frequency: $\text{freq}(v')$, $\text{freq}(e')$. By construction: $\sum_{v' \in V'} \text{freq}(v') = |V|$, $\sum_{e' \in E'} \text{freq}(e') = |E|$.

Folding Factor:

$$F = \frac{|V| + |E|}{|V'| + |E'|}. \quad (2)$$

A higher F indicates greater structural compression.

Folding Variance: Let $\mu_V = \frac{1}{|V'|} \sum_{v' \in V'} \text{freq}(v')$. The variance is:

$$\text{Var}_{\text{fold}}(V') = \frac{1}{|V'|} \sum_{v' \in V'} \text{freq}(v') (\text{freq}(v') - \mu_V)^2. \quad (3)$$

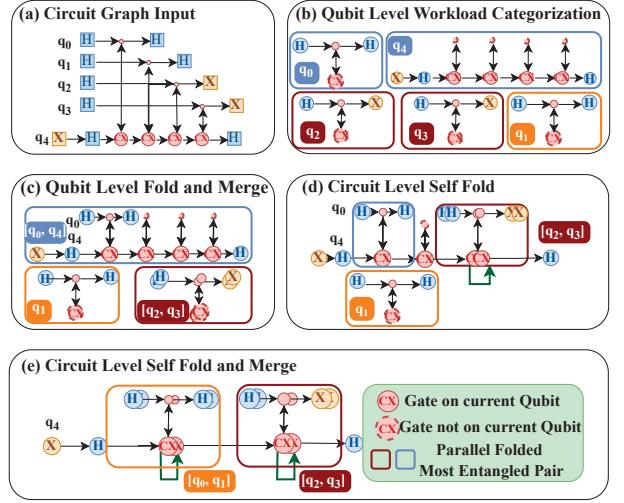


Fig. 4: Illustration of Circuit Folding: (a) Circuit-level graph. (b) Five qubit-level gate sequences are paired based on Maximum Entangled Pairs, highlighted in matching colors, and folded in parallel. (c) A common gate sequence between q_2 and q_3 is detected using LCCS function, merging the red-circled nodes. (d) The folded structure of $[q_2, q_3]$ triggers self-folding in q_4 , marked by a green edge. (e) The final meta-graph reduces node count from 19 to 11.

A lower $\text{Var}_{\text{fold}}(V')$ implies more uniform module sizes, simplifying partitioning. Together with F , this metric helps assess folding quality and avoid *under-folding* (low F) or *over-folding* (high variance dominated by few large modules), ensuring a compact and balanced meta-graph for efficient circuit cutting.

C. Framework Design: Circuit Folding

Figure 3 illustrates the CiFOLD framework using a 5-qubit example. First, the input circuit is translated into qubit-level sequence graphs. These per-qubit graphs are then analyzed by the folding algorithm (Algorithm 1, Figure 4) to detect repeated qubits with identical or similar workloads, CiFOLD consolidates redundant structures and captures essential patterns. The resulting meta-graph is subsequently “unfolded” to generate an optimized circuit, which is finally partitioned into smaller subcircuits (Algorithm 2, Figure 5). The partitioning

Algorithm 1 Parallel Folding of Qubit-Level Pattern Graphs

Input: Qubit-level sequence graphs $G_{q_0}, G_{q_1}, \dots, G_{q_n}$
Output: Folded meta-graph G'

- 1: **for all** $min_LCS_len \in [\max(\text{seq_len}) \dots 3]$ decreasing **do**
- 2: $G_{\text{pattern}} \leftarrow [G_{q_0}, G_{q_1}, \dots, G_{q_n}]$
- 3: **while** $|G_{\text{pattern}}| > 1$ **do**
- 4: $\text{MEP} \leftarrow \text{IdentifyMostEntangledPairs}(G_{\text{pattern}})$
- 5: **for all** $(G_{q_i}, G_{q_j}) \in \text{MEP}$ **in parallel do**
- 6: $[L_i, L_j] \leftarrow \text{LCCS}(G_{q_i}.\text{nodes}, G_{q_j}.\text{nodes}, min_LCS_len)$
- 7: **if** $[L_i, L_j] \neq \emptyset$ **then**
- 8: $G_{\text{merged}} \leftarrow \text{ComposeGraphs}(G_{q_i}, G_{q_j})$
- 9: **for all** $(v_i, v_j) \in [L_i, L_j]$ **do**
- 10: Transfer edges from v_j to v_i
- 11: Fold v_j into v_i
- 12: Replace G_{q_i}, G_{q_j} with G_{merged} in G_{pattern}
- 13: Compute F and $\text{Var}_{\text{fold}}(V')$ via Eq. (2), (3)
- 14: **if** no improvement in F and $\text{Var}_{\text{fold}}(V')$ **then**
- 15: **return** folded graph G'
- 16: **return** Folded graph G'

step reduces hardware demands and sampling overhead, allowing circuits to be executed efficiently on current quantum devices.

1) *Qubit-Level Gate Sequence:* Given the graph representation of input circuit $G = (V, E, w)$, the directed edges explicitly encode the inherent temporal ordering constraints imposed by quantum circuit operations. Leveraging these constraints, we simplify the task of detecting structural patterns by analyzing sequential gate ordering at the qubit level, thus avoiding the computational complexity associated with general subgraph isomorphism searches. Specifically, each qubit q_i yields a qubit-level sequence graph G_{q_i} , as illustrated in Figure 4(b). This extraction can be efficiently parallelized.

Given two qubit gate sequences, we employ dynamic programming to identify the Longest Consecutive Common Subsequence (LCCS). A dynamic programming table is initialized and systematically filled by comparing elements from both sequences, tracking the longest contiguous match. Only subsequences longer than a predefined threshold are retained, ensuring that the identified patterns are significant and suitable for use in the subsequent folding stage.

2) *Folding: Layered Frequent Pattern Discovery:* The folding process identifies repeated gate sequences, consolidating nodes and edges into a meta-graph. Each meta-node represents structurally equivalent groups from the original circuit, preserving gate parameters and connectivity.

Figure 4 demonstrates folding on a 5-qubit circuit. Algorithm 1 details the parallel folding method on qubit-level graphs $[G_{q_0}, G_{q_1}, \dots]$. Initially, a sequence-length threshold is set (Line 1), and pairs of qubits with the most shared gates (Most Entangled Pairs, MEPs) are identified (Lines 3–4). The Longest Common Consecutive Subsequence (LCCS) algorithm detects common sequences (Line 6), enabling node

Algorithm 2 Meta-Graph Guided Circuit Partitioning

Input: Circuit graph G , Folded graph G_f , qubit constraint q_{con}
Output: Partitioned subgraphs of the circuit graph

- 1: Initialize $\text{VisitedNodes} \leftarrow \emptyset$
- 2: $P_{\text{initial}} \leftarrow \text{ModuleFinding}(G, G_f)$ {Initial modules indexed by hash}
- 3: Construct graph G' where each subgraph in P_{initial} forms a single node
- 4: $G' \leftarrow \text{GreedyMerge}(q_{\text{con}}, P_{\text{initial}})$
- 5: $P \leftarrow \text{Refinement}(G', q_{\text{con}}, P_{\text{initial}})$
- 6: **return** Final partition P

Algorithm 3 Optimized Initial Partition via Edge Growth

Input: Circuit graph G , Folded meta-graph G_f
Output: Initial set of partitions P_{init}

- 1: Initialize $\text{visited_nodes} \leftarrow \emptyset, P_{\text{init}} \leftarrow \emptyset$
- 2: **while** $|\text{visited_nodes}| < |G.\text{nodes}|$ **do**
- 3: $v_{\text{freq}} \leftarrow \text{SelectMostFoldedNode}(G_f, \text{visited_nodes})$
- 4: **for all** $v \in v_{\text{freq}}.\text{folded_nodes}$ **in parallel do**
- 5: Initialize $\text{current_subgraph} \leftarrow \{v\}$
- 6: Identify candidate neighbors of v in G
- 7: **for all** neighbors n connected via edge e to v **do**
- 8: Add node n and edge e to current_subgraph
- 9: Compute $\text{hash} \leftarrow \text{WeisfeilerLehman}(\text{current_subgraph})$
- 10: Append current_subgraph to $P_{\text{init}}[\text{hash}]$
- 11: Update visited_nodes with nodes in current_subgraph
- 12: **return** P_{init}

collapsing (Lines 7–12). Edges are transferred to maintain structural integrity.

The folding operates iteratively, halving the number of graphs each step, thus requiring $\log_2 n$ iterations for n -qubit circuits. Each iteration maximizes the folding factor F (Eq. 2) while minimizing frequency variance (Eq. 3). The threshold is progressively lowered until no further improvement occurs.

3) *Unfolding: Modular Partitioning Search:* The meta-graph encapsulating repeated structures and connectivity for partitioning. Algorithm 3 details the initial partitioning strategy. An example is shown in Figure 5.

C1FOLD leverages Weisfeiler-Lehman (WL) hashing for efficiently detecting structurally similar subcircuits. Though WL hashing does not guarantee exact isomorphism, it sufficiently discriminates under quantum circuit constraints. Nodes are selected based on folding frequency, expanded using canonical edge ordering, and grouped using WL hashes.

Partitions are determined by selecting frequently occurring WL hashes. These partitions form supernodes for subsequent global partitioning, employing a METIS-inspired greedy merging and refinement strategy to minimize sampling overhead. Nodes near partition boundaries are exchangeable between partitions, further optimizing partition quality.

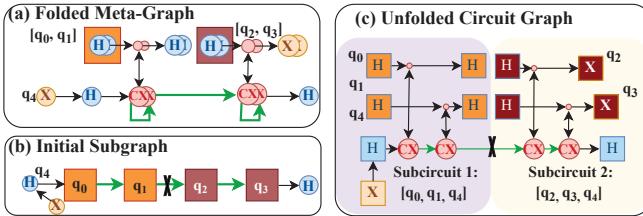


Fig. 5: **Illustration of Circuit Unfolding:** The yellow and red square represents the folded node (q_0, q_1) and (q_2, q_3) in (a). (b) illustrates the identified initial subgraphs, partitioned into two 3-qubit subcircuits. (c) displays the unfolded circuit, highlighting the cuts determined in (b).

V. EVALUATION

This section evaluates the effectiveness of the proposed CiFOLD framework against state-of-the-art circuit cutting methods across multiple dimensions: execution time, number of cuts, sampling overhead, and reconstruction fidelity.

A. Meta Graph Comparison

Figure 6 shows the graph representation of example 12 and 24-qubit circuit with its meta graph. The corresponding folding metrics in shown in I. These examples illustrate that a higher folding factor leads to greater reductions in graph size and complexity, while a higher folding variance indicates an uneven distribution of folding. For 12-qubit and 24-qubit circuit, they can converge to a similar meta-graph structure, differing only in frequency distributions with higher folding factor for 24-qubit circuit. Each folded node in the meta-graph aggregates multiple nodes from the original circuit, with its frequency count reflecting the extent of consolidation. Additionally, nodes of the same color correspond to identical gate types. This folding approach effectively simplifies large-scale quantum circuits while preserving their essential structure, facilitating more efficient partitioning and execution under hardware constraints.

TABLE I: Folding Factor and Folding Variance

Circuit Type	Qubit Count	F	$\text{Var}_{\text{fold}}(V')$
QAOA	12	7.32	24.00
BV	12	6.64	24.45
QAOA	24	18.08	44.62
BV	24	13.50	91.08

B. Workload, Implementation and Settings

Benchmark Circuits: We evaluate and compare CiFOLD using five quantum algorithms commonly employed in previous circuit-cutting studies [1], [2], [4]: (1) **QAOA:** Hardware-efficient ansatz for combinatorial optimization [32]; (2) **BV:** Bernstein-Vazirani algorithm for exponential oracle-query speedup [13]; (3) **GHZ:** Entangled state preparation for quantum communication [36]; (4) **W State:** Robust multipartite entanglement; (5) **Ising:** 2-local Hamiltonian simulation for optimization [30], [37].

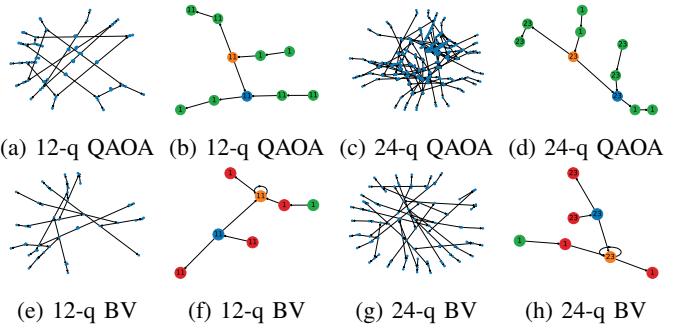


Fig. 6: Comparison of Original Graph and Meta Graph: Each node in meta graph consolidates multiple nodes in original graph with frequency indicated. Additionally, nodes of the same color correspond to same gate types.

Implementation and Experiment Settings: We implement the CiFOLD with the following software: Python 3.10 [38], IBM Qiskit 1.02 [39], Qiskit-Addon-Cutting 0.9.0 [25], and Networkx 3.3 [40]. The classical components of CiFOLD are executed on an AMD Ryzen 7 6800H processor running at 3.2 GHz. The quantum processors used are IBM Qiskit Emulators, using hardware calibration data from real machines including Auckland, Brisbane, CairoV2, Cusco, WashingtonV2, SydneyV2 and MontrealV2.

Baselines: We compare CiFOLD against three state-of-the-art circuit-cutting frameworks: **Qiskit-Addon-Cutting** [25] utilizes a Dijkstra-based best-first search algorithm supporting both gate and wire cuts. **CutQC** [2] employs solver-based optimization restricted to wire cuts, guaranteeing optimal solutions for **wire-only** partitions. However, its factorial runtime growth significantly limits scalability. To manage this, a 300-second time limit is enforced, after which the best available solution (or none, if unresolved) is returned. **FitCut** [9], another **wire-only** method, combines Louvain community detection with greedy merging for efficient partitioning.

C. Evaluation Metrics

Relative Fidelity: We assess fidelity through an end-to-end pipeline as shown on Figure 3. We compare the reconstructed expectation value for computational Z basis from partitioned circuits against values obtained by direct, unpartitioned execution on same noisy emulators.

Execution Time: Execution time reflects the overhead of partitioning, with faster runtimes indicating better scalability for large circuits. A red cross marks cases exceeding the 300-second timeout.

Number of Cuts: The number of cuts directly influences the exponential classical post-processing overhead ($O(4^n)$ per wire cut). Therefore, fewer cuts indicate more efficient solutions and reduced post-processing costs.

Sampling Overhead: Sampling overhead measures how many times subcircuits must be executed to achieve accurate circuit reconstruction. Hybrid approaches integrating gate and wire cuts (such as CiFOLD and Qiskit-Addon-Cutting) typically

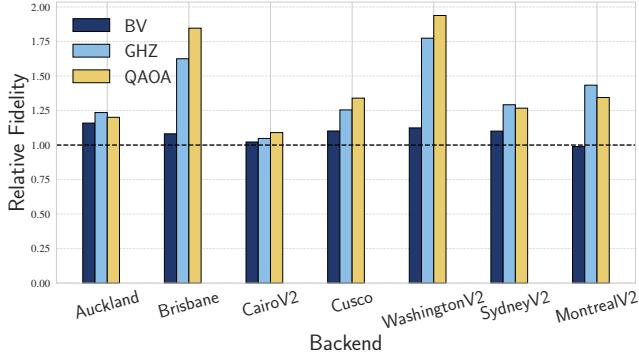


Fig. 7: Relative fidelity of CiFOLD compared to direct execution for three 20-qubit circuits, computed as the normalized difference from uncut execution. Dashed line denotes baseline fidelity (uncut execution).

outperform frameworks relying exclusively on wire cuts (e.g., CutQC, FitCut).

D. Performance Analysis

Relative Fidelity: In Figure 7, we present the relative fidelity of reconstructed expectation values for a 20-qubit circuit with 10-qubit constraint across BV, QAOA, and GHZ benchmarks. The evaluation is conducted on seven IBM backend emulators to account for different noise profiles. The average relative fidelity improvement factors are 19.8%, 51.8%, 5.3%, 23.2%, 61.2%, and 22.0% for Auckland, Brisbane, CairoV2, Cusco, WashingtonV2, SydneyV2, and MontrealV2, respectively. Therefore, CiFOLD demonstrates an improvement in fidelity due to reduced subcircuit width and depth, consistent with findings from previous studies [2], [4], [23].

Execution Time: As shown in Fig. 8 (left), the dashed line marks the 1-second threshold, and a red cross indicates a 300-second timeout. CiFOLD consistently maintains runtimes near or below this mark across all benchmarks, leveraging parallelized folding and unfolding processes. FitCut consistently demonstrates the lowest runtimes (around 0.1s). This is because it is exclusively reliant on wire cuts and simplifies graph complexity compared to the hybrid gate and wire cut approaches like CiFOLD. Without supporting gate cuts, FitCut leads to a larger number of cuts and a significant increase in sampling overhead as shown in Fig. 8(center and right figures).

Besides, CiFOLD outperforms Qiskit-Addon-Cutting significantly. For example, the average execution times are 1.13s vs 12.97s (Ising), 0.30s vs 36.91s (GHZ), 1.19s vs 25.76s (W-State), 0.41s vs 9.65s (BV), and 0.68s vs 8.62s (QAOA) for CiFOLD and Qiskit-Addon-Cutting, respectively. This represents an average 94.7% reduction. CutQC matches CiFOLD performance for smaller circuits (typically under 1s), but suffers prohibitive factorial runtime growth. For example, CutQC fails to return a cutting solution for many 60-qubit circuits due to 300s API timeouts. Specifically, with 60-qubit Ising, Qiskit-Addon-Cutting and CiFOLD return a solution in 37.3s and 1.83, but CutQC fails with 300s timeout. Notably,

CutQC’s runtime grows from 2.31s to 300s(timeout) for a 10-qubit constraint with 42- and 66-qubit Ising circuits.

Number of Cuts: Reducing the number of cuts directly decreases classical post-processing overhead. As shown in Fig. 8 (center), CiFOLD consistently yields fewer or comparable cuts than other methods. A red cross indicates that, within the 300-second timeout, either the best result was found or no result was returned.

For 24-qubit BV, QAOA, GHZ and Ising circuits, all evaluated methods achieve same cut counts due to the relatively low complexity. Notably, in the 24-qubit W-State benchmark, CiFOLD and Qiskit-Addon-Cutting only require 4 and 5 cuts respectively compared to 6 cuts for CutQC and FitCut, attributable to their hybrid gate-wire cutting capability.

Performance differences become more pronounced as circuit size increases. Qiskit-Addon-Cutting lacks sufficient optimization, resulting in significantly higher cut counts—up to 50 cuts for BV and 12 cuts for Ising—compared to CiFOLD’s 5 and 6 cuts. Despite CutQC’s optimality guarantee for wire-only cutting, CiFOLD achieves fewer cuts for 66-qubit Ising benchmark with 10-qubit constraint where CutQC requires 7 cuts and CiFOLD identifies 6. CiFOLD shows a clear advantage in the W-State benchmark, where CutQC times out on the 60-qubit circuit and yields 14 cuts for the 48-qubit case under a 15-qubit constraint. In contrast, CiFOLD efficiently identifies viable solutions with just 6 cuts. These outcomes emphasize the precision of CiFOLD in identifying cut locations, consistently minimizing the number of cuts across various circuit architectures while also significantly reducing execution time compared to solver-based method.

Overall, CiFOLD achieves an average reduction of 31.6% in the number of cuts (average of 3.5 cuts), compared to Qiskit-Addon-Cutting (7.8 cuts, 55.2% reduction), CutQC (3.9 cuts, 10.0% reduction), and FitCut (5.0 cuts, 29.5% reduction). Specifically, CiFOLD substantially outperforms Qiskit-Addon-Cutting due to its limited optimization, moderately improves upon CutQC despite its theoretical wire-only optimality guarantees, and consistently surpasses FitCut, whose greedy merging approach tends to become trapped in suboptimal local minima. Extending to larger circuits beyond 60 qubits, where solver-based methods such as CutQC encounter timeouts, we expect CiFOLD to deliver even greater performance gains.

Sampling Overhead: Figure 8(right figures) present the sampling overhead (computed by Eq.1). It directly quantifies the computational cost of circuit-cutting methods, growing exponentially with each additional cut. While strongly correlated with the number of cuts, sampling overhead also depends on the type of cuts chosen, giving hybrid gate-and-wire approaches such as CiFOLD and Qiskit-Addon-Cutting a notable advantage. For the 24-qubit QAOA circuit, although CutQC and FitCut match CiFOLD’s cut count, their wire-only sampling overhead of 256 vastly exceeds CiFOLD’s hybrid overhead of 81. Similarly, for the 60-qubit QAOA benchmark, six wire cuts incur a sampling overhead of 1.68×10^7 , whereas CiFOLD’s five hybrid cuts require only 5.90×10^4 , further highlighting the advantage of hybrid gate-wire cutting.



Fig. 8: Performance Benchmarks. The x-axis labels follow the format $\{\text{size}\}_{\{\text{constraint}\}}$, covering circuit sizes from 24 to 66 qubits under 10-qubit and 15-qubit constraints. The dashed line represents the 1-second threshold and red cross indicates a 300 second timeout.

Overall, CiFOLD’s 31.6% average reduction in cuts corresponds to an average sampling overhead of 8.88×10^7 , compared to Qiskit’s 1.28×10^{46} , CutQC’s 1.94×10^{13} , and FitCut’s 3.00×10^{22} , with an overall reduction factor of 3.55×10^9 (excluding CutQC Timeouts). It underscores the advantage of hybrid gate–wire circuit-cutting strategies.

VI. CONCLUSION

In this work, we proposed CiFOLD, a hybrid gate- and wire-cutting framework that leverages modular and repetitive structures common in quantum circuits. Utilizing graph-based folding and unfolding strategies, CiFOLD efficiently reduces circuit partitioning complexity and sampling overhead under

practical hardware constraints. We introduced metrics such as the *folding factor* and *weighted node folding variance* to quantify and optimize the balance between circuit compression and reconstruction accuracy.

Empirical evaluations demonstrate that CiFOLD maintains high measurement fidelity while outperforming existing circuit-cutting methods in terms of scalability, overhead, and runtime efficiency. Future work includes integrating advanced cutting protocols (e.g., parallel cuts and classical communication) and developing adaptive strategies for dynamic circuit partitioning in distributed quantum systems. Ultimately, CiFOLD presents a promising direction for scalable and robust quantum computing within realistic hardware environments.

REFERENCES

[1] W. Tang and M. Martonosi, "Tensorqc: Towards scalable distributed quantum computing via tensor networks," 2025. [Online]. Available: <https://arxiv.org/abs/2502.03445>

[2] W. Tang, T. Tomesh, M. Suchara, J. Larson, and M. Martonosi, "Cutqc: using small quantum computers for large quantum circuit evaluations," in *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '21. ACM, Apr. 2021. [Online]. Available: <http://dx.doi.org/10.1145/3445814.3446758>

[3] A. Lowe, M. Medvidović, A. Hayes, L. J. O'Riordan, T. R. Bromley, J. M. Arrazola, and N. Killoran, "Fast quantum circuit cutting with randomized measurements," *Quantum*, vol. 7, p. 934, 2023.

[4] X. Ren, M. Zhang, and A. Barbalace, "A hardware-aware gate cutting framework for practical quantum circuit knitting," in *Proceedings of the 2024 IEEE/ACM International Conference on Computer-Aided Design (ICCAD '24)*. IEEE/ACM, 2024.

[5] C. Utrecht, M. Periyasamy, S. Rietsch, D. D. Scherer, A. Plinge, and C. Mutschler, "Cutting multi-control quantum gates with zx calculus," *Quantum*, vol. 7, p. 1147, 2023.

[6] L. Schmitt, C. Piveteau, and D. Sutter, "Cutting circuits with multiple two-qubit unitaries," 2024.

[7] S. Brandhofer, I. Polian, and K. Krsulich, "Optimal partitioning of quantum circuits using gate cuts and wire cuts," *IEEE Transactions on Quantum Engineering*, 2023.

[8] A. Pawar, Y. Li, Z. Mo, Y. Guo, X. Tang, Y. Zhang, and J. Yang, "Qrc: Evaluating large quantum circuits on small quantum computers through integrated qubit reuse and circuit cutting," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4*, 2024, pp. 236–251.

[9] S. Kan, Z. Du, M. Palma, S. A. Stein, C. Liu, W. Wei, J. Chen, A. Li, and Y. Mao, "Scalable circuit cutting and scheduling in a resource-constrained and distributed quantum system," *2024 IEEE Quantum Computing and Engineering (QCE)*, 2024.

[10] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 1996, pp. 212–219.

[11] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge university press, 2010.

[12] S. A. Cuccaro, T. G. Draper, S. A. Kutin, and D. P. Moulton, "A new quantum ripple-carry addition circuit," *arXiv preprint quant-ph/0410184*, 2004.

[13] E. Bernstein and U. Vazirani, "Quantum complexity theory," in *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, 1993, pp. 11–20.

[14] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th annual symposium on foundations of computer science*. Ieee, 1994, pp. 124–134.

[15] K. Mitarai and K. Fujii, "Constructing a virtual two-qubit gate by sampling single-qubit operations," *New Journal of Physics*, vol. 23, no. 2, p. 023021, 2021.

[16] ——, "Overhead for simulating a non-local channel with local channels by quasiprobability sampling," *Quantum*, vol. 5, p. 388, Jan. 2021. [Online]. Available: <https://doi.org/10.22331/q-2021-01-28-388>

[17] L. Brenner, C. Piveteau, and D. Sutter, "Optimal wire cutting with classical communication," 2023.

[18] C. Piveteau and D. Sutter, "Circuit knitting with classical communication," *IEEE Transactions on Information Theory*, 2023.

[19] H. Harada, K. Wada, and N. Yamamoto, "Doubly optimal parallel wire cutting without ancilla qubits," *PRX Quantum*, vol. 5, p. 040308, Oct 2024. [Online]. Available: <https://link.aps.org/doi/10.1103/PRXQuantum.5.040308>

[20] E. Pednault, "An alternative approach to optimal wire cutting without ancilla qubits," 2023.

[21] D. Chen, B. Baheri, V. Chaudhary, Q. Guan, N. Xie, and S. Xu, "Approximate quantum circuit reconstruction," in *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, 2022, pp. 509–515.

[22] D. T. Chen, E. H. Hansen, X. Li, V. Kulkarni, V. Chaudhary, B. Ren, Q. Guan, S. Kuppannagari, J. Liu, and S. Xu, "Efficient quantum circuit cutting by neglecting basis elements," in *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2023, pp. 517–523.

[23] G. Gentinetta, F. Metz, and G. Carleo, "Overhead-constrained circuit knitting for variational quantum dynamics," *Quantum*, vol. 8, p. 1296, Mar. 2024. [Online]. Available: <https://doi.org/10.22331/q-2024-03-21-1296>

[24] Z. Li, M. Guo, M. Barad, W. Tang, E. Z. Zhang, and Y. Huang, "A case for quantum circuit cutting for nisq applications: Impact of topology, determinism, and sparsity," 2024. [Online]. Available: <https://arxiv.org/abs/2412.17929>

[25] A. M. Brańczyk, A. Carrera Vazquez, D. J. Egger, B. Fuller, J. Gacon, J. R. Garrison, J. R. Glick, C. Johnson, S. Joshi, E. Pednault, C. D. Pemmaraju, P. Rivero, I. Shehzad, and S. Woerner, "Qiskit addon: circuit cutting," <https://github.com/Qiskit/qiskit-addon-cutting>, 2024.

[26] P. Li, J. Liu, A. Gonzales, Z. H. Saleem, H. Zhou, and P. Hovland, "Qutracer: Mitigating quantum gate and measurement errors by tracing subsets of qubits," 2024. [Online]. Available: <https://arxiv.org/abs/2404.19712>

[27] C. Utrecht, L. S. Herzog, D. D. Scherer, M. Periyasamy, S. Rietsch, A. Plinge, and C. Mutschler, "Optimal joint cutting of two-qubit rotation gates," *Physical Review A*, vol. 109, no. 5, p. 052440, 2024.

[28] A. Li, S. Stein, S. Krishnamoorthy, and J. Ang, "Qasmbench: A low-level quantum benchmark suite for nisq evaluation and simulation," *ACM Transactions on Quantum Computing*, vol. 4, no. 2, pp. 1–26, 2023.

[29] T. Tomesh, P. Gokhale, V. Omole, G. S. Ravi, K. N. Smith, J. Viszlai, X.-C. Wu, N. Hardavellas, M. R. Martonosi, and F. T. Chong, "Supermarq: A scalable quantum benchmark suite," in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 587–603.

[30] S. Suzuki, J.-i. Inoue, and B. K. Chakrabarti, *Quantum Ising phases and transitions in transverse Ising models*. Springer, 2012, vol. 862.

[31] S. A. Stein, B. Baheri, D. Chen, Y. Mao, Q. Guan, A. Li, S. Xu, and C. Ding, "Quclassi: A hybrid deep neural network architecture based on quantum state fidelity," in *Proceedings of Machine Learning and Systems*, D. Marculescu, Y. Chi, and C. Wu, Eds., vol. 4, 2022, pp. 251–264.

[32] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets," *Nature*, vol. 549, no. 7671, p. 242–246, Sep. 2017. [Online]. Available: <http://dx.doi.org/10.1038/nature23879>

[33] J. Tilly, H. Chen, S. Cao, D. Picozzi, K. Setia, Y. Li, E. Grant, L. Wossnig, I. Runger, G. H. Booth *et al.*, "The variational quantum eigensolver: a review of methods and best practices," *Physics Reports*, vol. 986, pp. 1–128, 2022.

[34] X. Yan and J. Han, "gspan: graph-based substructure pattern mining," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings*, 2002, pp. 721–724.

[35] C. Borgelt and M. R. Berthold, "Mining molecular fragments: Finding relevant substructures of molecules," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings*. IEEE, 2002, pp. 51–58.

[36] D. M. Greenberger, M. A. Horne, and A. Zeilinger, "Going beyond bell's theorem," *Bell's theorem, quantum theory and conceptions of the universe*, pp. 69–72, 1989.

[37] H. Labuhn, D. Barredo, S. Ravets, S. De Léséleuc, T. Macrì, T. Lahaye, and A. Browaeys, "Tunable two-dimensional arrays of single rydberg atoms for realizing quantum ising models," *Nature*, vol. 534, no. 7609, pp. 667–670, 2016.

[38] Python Software Foundation, "Python Language Reference, version 3.10," 2021. [Online]. Available: <https://www.python.org/>

[39] "IBM Quantum Computing — Qiskit — ibm.com," <https://www.ibm.com/quantum/qiskit>, [Accessed 21-04-2024].

[40] A. A. Hagberg, D. A. Schult, and P. J. Swart, "NetworkX: Network Analysis in Python," 2024, version 3.3. [Online]. Available: <https://networkx.github.io/>