Peephole Optimization for Quantum Approximate Synthesis

***Abstract*—Peephole optimization of quantum circuits provides a method of leveraging standard circuit synthesis approaches into scalable quantum circuit optimization. One application of this technique partitions the entire circuit into a series of peepholes and produces multiple approximations of each partitioned subcircuit. One approximation of each subcircuit is then selected to form optimized result circuits. We propose a series of improvements to the final phase of this architecture, which include the addition of error awareness and a better method of approximating the correctness of the result. We evaluated these proposed improvements on a set of benchmark circuits using the IBMQ FakeWashington simulator. The results demonstrate that our best-performing method provides an average reduction in Total Variational Distance (TVD) and Jensen-Shannon Divergence (JSD) of 15.0% and 13.1%, respectively, compared with the Qiskit optimizer. This also constitutes an improvement in TVD of 11.4% and JSD of 9.0% over existing solutions.**

I. Introduction

Peephole optimization of quantum circuits is a very effective and scalable optimization technique which selects classically-tractable sections (peepholes) of a quantum circuit and optimizes each section. This allows optimization techniques with poor scaling, such as resynthesis, to be applied to large circuits, although only to small sections at a time.

Full-circuit peephole optimization methods, such as [1], [2], partition whole quantum circuits into peepholes, and then perform resynthesis on each partitioned component. The resulting component circuits are then reassembled into an optimized full circuit, taking the better of each pair of subcircuits (the original or the resynthesized version). A variant of this technique, Quest [3], generates multiple approximations of each partition and attempts to create a set of result circuits which can more closely match the ideal output than the original circuit when executed on noisy hardware. This is accomplished by adding an additional step to the synthesis process, which we call "recombination", shown in Figure 1. As the recombination step is responsible selecting partition approximations to produce optimal circuits, it has a substantial effect on the quality of the result. Thus, improving the recombination step would produce significant performance improvements.

The recombination technique employed in Quest is a dual annealer which explores the set of possible subcircuit combinations. The chosen objective function is composed of three main metrics: (1) one which ensures that process distance between the approximation and the original circuit is within some acceptable range (by default 0.1); (2) a complexity reduction metric which reduces the number of CNOT gates, which is meant to minimize the effect of hardware error on the circuit; and (3) a differentiation metric which encourages the selection of result circuits that are different than the ones already selected.

This method produces good results for many applications, but it has some significant limitations. 1) Although limiting the approximation error of each result circuit while minimizing the number of multi-qubit gates is theoretically sound, in practice it leaves much to be desired. In addition to introducing another parameter to consider, the approach fails to consider other sources of error, such as thermal noise and interactions with the environment, which more strongly correlated to circuit depth than CNOT count. 2) In addition, while the sum of partition process distances is proven to provide an upper bound on the overall process distance, estimating circuit performance this way does not give any consideration to the interactions between partitions. For example, it may turn out that a small error in one partition becomes much larger when propagated to the next partition, or that a large error in one partition mostly cancels out with another error in the next. 3) Iteratively selecting partitions has the effect of producing better approximations at the beginning of the process, and significantly worse approximations later on as more circuits exist to compare new circuits against.

Additionally, while this method has been demonstrated to perform well in more favorable conditions (smaller circuits or uniform noise, fully connected hardware), testing reveals that when circuits are mapped onto hardware with more complex errors and limited connectivity, performance is substantially degraded.

We propose three new recombination techniques to address these limitations, in addition to making several smaller optimizations to the original method and some changes to the original flow. To address the first limitation, we propose an error aware circuit fidelity evaluation, which combines the apparently opposed objectives of retaining circuit functionality and reducing CNOT count while also accounting for other sources of error. To address the second point, we implement a cascaded error estimation method, which considers partition pairs rather than individual partitions. This allows the method to account for the error which happens as a result of interactions between partitions. To address the third point, we implement a population-based annealing approach, which performs annealing on multiple candidates at once, and provides all candidates to the objective function for evaluation.

To evaluate the proposed techniques, we created four recombination configurations, each of which implements one or more of the proposed techniques. We have also included the recombination method used in Quest, as well as an improved variant of that method. These configurations were evaluated by mapping a series of test circuits to the IBMQ Washington computer, running them through the approximation process, optimizing the results with Qiksit, and simulating using the

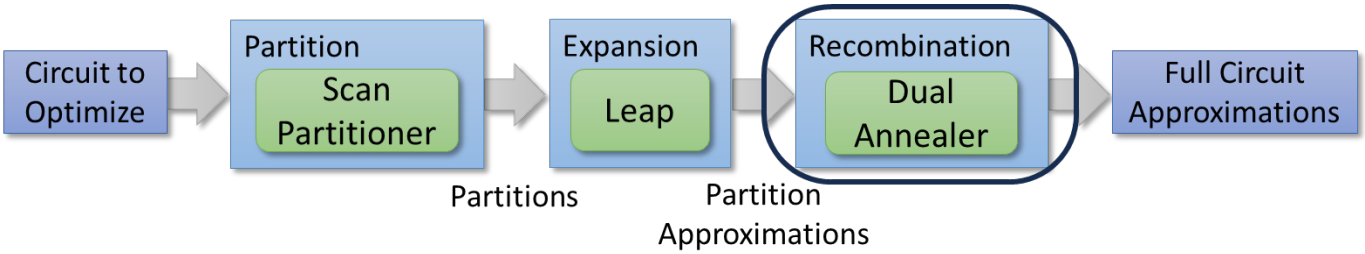

Fig. 1: Basic structure of the Quest algorithm, with three phases. Partitioning splits the circuit, expansion approximates each partition, and recombination puts approximations together to produce one or more noise resilient approximations. Recombination (circled) is the focus of this work.

FakeWashington backend included in Qiskit. The most promising configuration, the population-based method with error awareness, achieves a reduction in Total Variation Distance (TVD) from the ideal result of 18.2% and in Jensen-Shannon Divergence (JSD) of 15.8% when compared with the original mapped circuit. When compared with the result from the Quest method, our best method reduces TVD by 11.4% and JSD by 9.0%. This method also reduces multi-qubit gate count by an average of 37.1% from the baseline and 16.9% over the Quest method.

This paper is organized as follows: Section II describes our proposed techniques and recombination configurations. Section III discusses how the recombiner configurations were evaluated along with results. Section V discusses future research direction and concludes the article.

II. Proposed Methods

Algorithm 1 Basic structure of annealing objective function.

$P_i \leftarrow$ Original partitions
$P \leftarrow$ Partitions to evaluate
$S \leftarrow$ The set of existing partitions
$\epsilon \leftarrow$ Approximation threshold
if $P \in S$ **then**
    return 2.2
else if $\langle P|P_i\rangle_{HS} > \epsilon$ **then**
    return $\langle P_i|P\rangle_{HS} - \epsilon + 1.1$
else
    $t \leftarrow 0$
    for all $s \in S$ **do**
        $t \mathrel{+}= \langle P|s\rangle_{HS} \leq \max\left(\langle P_i|P\rangle_{HS}, \langle P_i|s\rangle_{HS}\right)$
    end for
    $t \mathrel{/}= |S|$
    $g = \text{CNOT}(P)$ / $\text{CNOT}(P_i)$
    return $w \times g + (1 - w) \times t$
end if

Our three proposed methods each rework a different part of the desired objective function. The cascaded error estimation improves the accuracy of the approximation limitation, while the error aware fidelity evaluation combines the approximation limitation and complexity reduction steps to produce an estimation of fidelity on the target hardware. Finally, the population-based annealer allows the objective function to be optimized over all result circuits at once, rather than iteratively producing single circuits, which ensures all circuits are equally affected by the differentiation metric.

In addition to the three proposed methods, we have implemented several smaller changes to the Quest objective function and implemented them in our own methods where relevant. First, we modify the approximation limitation to produce a gradient based on the amount of excess distance between the approximate and exact circuits rather than returning a constant value. This allows the annealer to explore the search space significantly more efficiently, and tends to allow access to formerly inaccessible regions of the search space. We also corrected a small but significant error in the differentiation metric which caused a tendency for results to resemble the initial exact circuit. The basic structure of this method is shown in Algorithm 1, which acts as the baseline configuration for all of our approaches.

A. Cascaded Error Estimation

The cascaded error estimation metric provides a more accurate estimate of the approximation error of a circuit by cascading the unitaries for adjacent partitions and calculating the process distance between that result and the same pair of partitions in the original circuit rather than comparing individual partitions. To facilitate this calculation, we construct a graph of the partition order, where nodes represent partitions and edges represent the flow of information between them, in the form of qubits. In order to evaluate a partition, we take the average of the distances for each pairing of a partition and its immediate neighbors on the partition graph. Each edge connecting a pair partitions proportionally increases the weight of that pair. To evaluate a circuit, we simply sum the scores of each partition composing the circuit.

B. Error Aware Fidelity Evaluation

The introduction of an error aware fidelity evaluation provides a significant structural improvement for our objective function by combining two seemingly opposing metrics and making the minimum accuracy parameter obsolete. We implement this objective function by calculating the probability

TABLE I: Description of the six different configurations.

Configuration	Basic Changes	Cascaded Error	Error Awareness	Population-Based
Quest				
Basic	X			
Basic w/Err	X		X	
Pop.	X			X
Pop. w/Err	X		X	X
Cascade	X	X		

density matrix of each partition in the initial circuit to use as the baseline. We then calculate the probability density matrix for each approximation running in an error simulation without readout error. Thus, rather than finding the process distance between the unitaries of the approximations and the exact circuit, we calculate the process distance between the ideal density matrix and one which results from the error simulation. We use the average of the distances of all circuit partitions to as both the fidelity estimate, which we use as the complexity reduction metric in place of reducing multi-qubit gates.

C. Population-Based Annealing

The Quest recombination approach performs the recombination algorithm once for each desired result circuit, adding each result to a list of prior results. The prior results are then used in the differentiation metric to score new circuits. However, this means that the first circuit produced does not account for any other circuits, while the last circuit is expected to be differentiated from all other circuits. Thus, a set of well-distributed approximate circuits which average to cancel out hardware error, this approach allows earlier circuits to have minimal CNOT count, while later circuits tend to become increasingly large. To address these concerns, we propose a population-based annealer, which performs annealing on each member of a population of candidate solutions simultaneously. This allows all solutions to be equally influenced by the differentiation metric. In order to implement this metric, we have modified several sections of an existing dual annealing implementation [4]. Namely, the main loop of the annealer now updates all solutions in each timestep, and saves a set of results when reannealing, rather individual results. We also added an argument to the objective function which contains all solutions except the solution to be evaluated, to enable the implementation of the differentiation metric. In addition to these changes, we also modify the objective function to allow duplicate results, as the improved differentiation behavior should allows the annealer to decide if duplicates are desirable.

D. Configurations

Aside from the cascaded error estimation and the error aware fidelity evaluation (which affect the same parts of the evaluation) the proposed improvements can be applied in tandem. As a result, we have produced five separate candidate configurations in addition to the Quest method, which are shown in Table I. The first of these configurations is our improved version of the Quest method, which is the basic structure on which our other configurations are built. The next two are one which employs the population-based approach and one which uses the cascaded error estimation, each with no additional changes. The final two are error aware variants of the improved Quest method and the population-based approach.

TABLE II: Benchmark circuits ysed to evaluate recombination methods.

Circuit	Description	Qubit Count	CNOT Count
Adder	Quantum adder	4 9	24 98
HLF	Hidden linear function circuit	5 10	14 56
Multiplier	Quantum multiplier	5 10	20 163
QAOA	Quantum approximate optimization algorithm	5 10	42 85
QFT	Quantum Fourier transform circuit	5 10	33 216
TFIM	Transverse-field Ising model simulation	4 8	12 56
XY	XY quantum Heisenberg model	4 8	12 56

III. Results

The six recombiners were implemented in the BQSKit quantum synthesis library [5] and evaluated by running the Quest pipeline and applying each recombiner to the same set of approximations for each circuit. The benchmark circuits are provided in Table II, along with a brief description of each circuit and the qubit and CNOT gate counts of each circuit. All six recombiners were applied to each testbench circuit and the resulting circuits, along with the initial hardware mapped circuit, were optimized with Qiskit with all optimizations on and simulated on the IBMQ FakeWashington backend with 1024 shots for each circuit. The mapped circuit was also run in an ideal simulator with optimizations disabled at 8192 shots. The Total Variational Distance and Jensen-Shannon Divergence of each combined set of results circuits in comparison with the ideal results were calculated. The results are presented in Figures 2 and 3.

The results demonstrate that while the Quest method performs well on a few circuits, most notably QFT 5, performance is generally worse than the other methods tested. Similarly, both the Basic and Basic with Error Awareness methods do not demonstrate impressive performance for any of the benchmark circuits. The Basic with Error configuration also performs significantly worse than any other configuration on HLF 5. The population-based method demonstrates some encouraging results, producing the best results out of all methods on HLF 5, and does not have any particularly poor benchmarks. However, the two best configurations by far are the population-based method with error awareness and the cascaded error estimation. The population-based method with error awareness gives at least some improvement on almost all test circuits, the only exception being Adder 9, which all configurations perform somewhat poorly on. However, even on Adder 9, Population with Error is among the better performing circuits. The

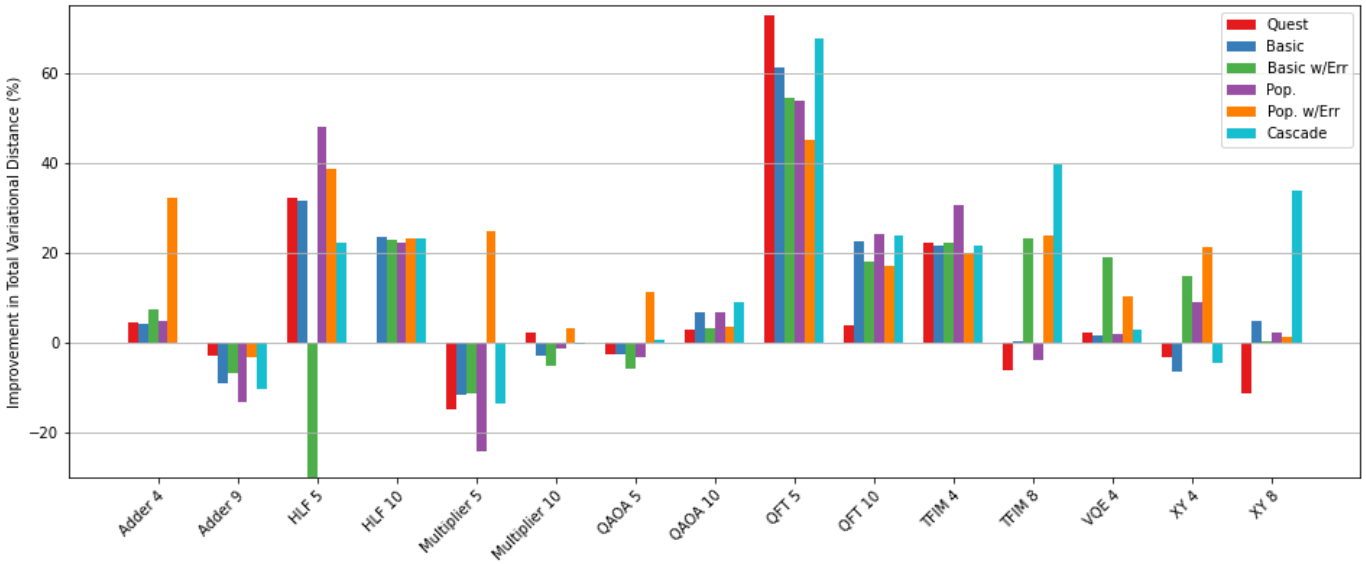

Fig. 2: Improvement in Total Variational Distance across all recombination configurations for all benchmark circuits. Not shown is the performance of the basic method with error awareness for HLF 5, which is -92.1%.

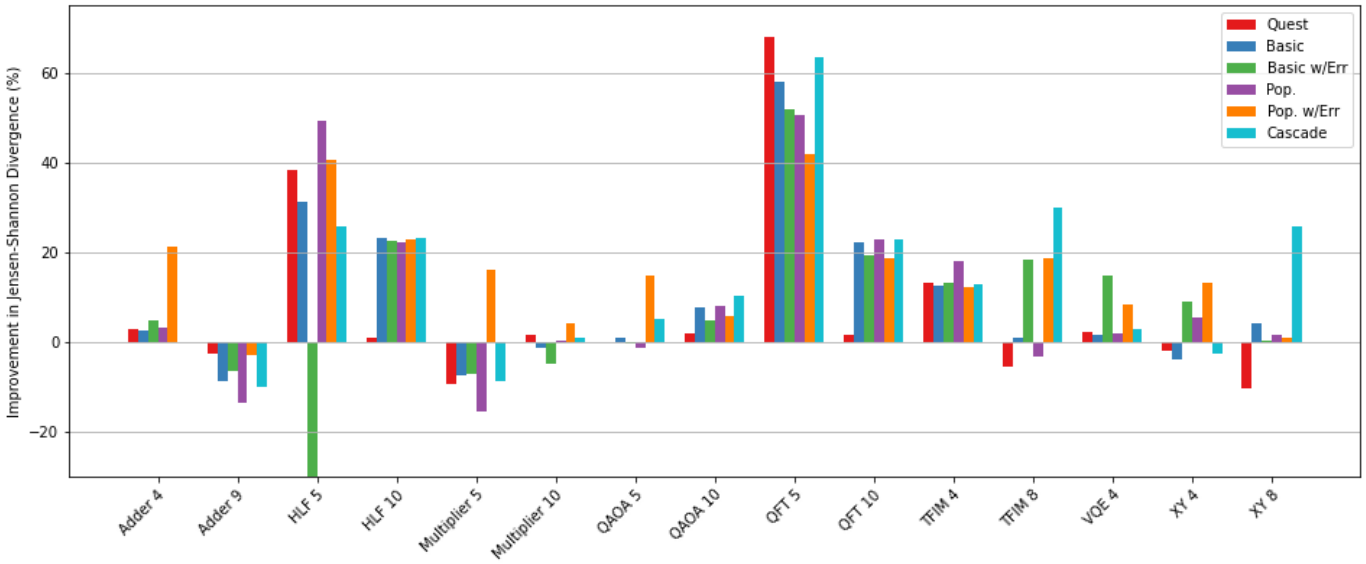

Fig. 3: Improvement in Jensen-Shannon Divergence across all recombination configurations for all benchmark circuits. Not shown is the performance of the basic method with error awareness for HLF 5, which is -71.1%.

Cascade approach does not provide the best performance on every circuit, but it is among the better performing approaches for most benchmarks, and performs far better than any other circuit on the TFIM 8 and XY 8 benchmarks.

The average reduction in the number of CNOT gates for each set of results was also calculated with respect to the exact circuit, the results for which are shown in Figure 4. The figure shows that all recombiners generally reduce the number of CNOT gates in the original circuit, in some cases by up to 80%, although there is still significant variation between recombiners. For example, the Quest approach generally offers the lowest reduction in CNOTs, often not reducing CNOT count at all. The Basic, Cascade and Population approaches generally offer similar reductions in CNOT count, with Cascade being the lowest and Population being the highest. The methods which stand out the most are the two error aware methods, which often produce considerably more reductions than the other methods. However, in several cases, the Population with Error method actually produces considerably less reduction in CNOT count than other circuits.

Table III shows the average performance improvement for each method on each benchmark circuit in terms of TVD, JSD,

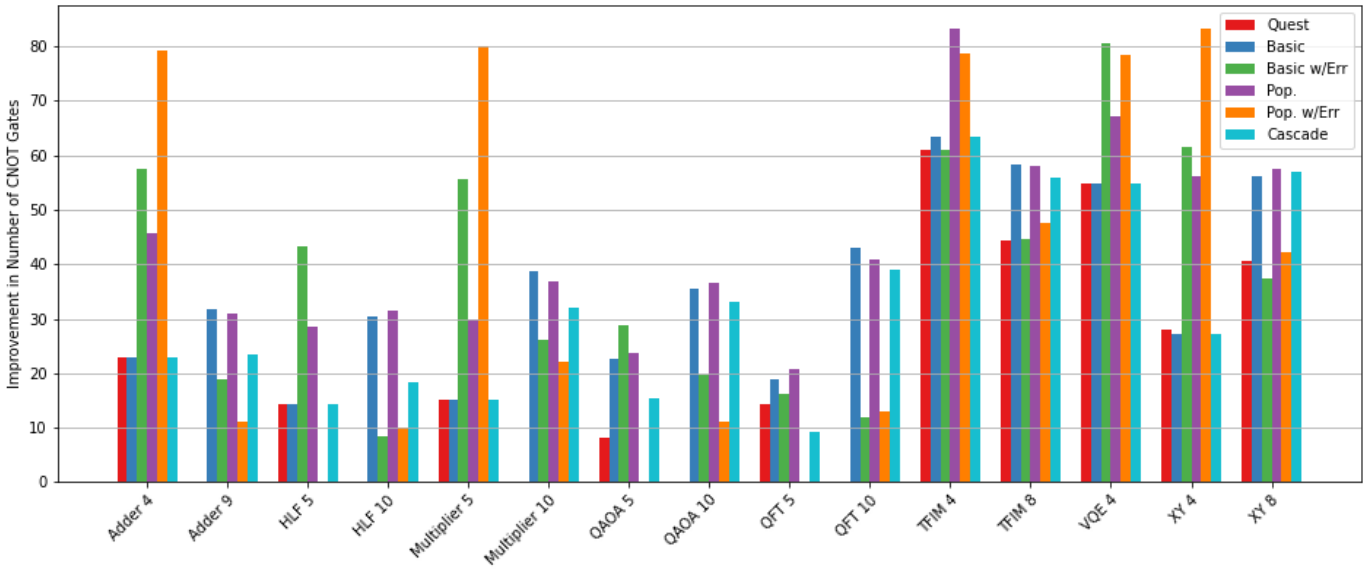

Fig. 4: Improvement in number of CNOT gates across all recombination configurations for all benchmark circuits. Not shown is the performance of the basic method with error awareness for HLF 5, which is -71.1%.

TABLE III: Performance improvement of recombination methods over the original circuit.

		Proposed Methods				
Metric	Quest [3]	Basic	Basic w/Err	Pop.	Pop. w/Err	Cascade
TVD	6.8%	9.8%	4.4%	10.5%	18.2%	14.4%
JSD	6.8%	9.7%	4.7%	10.1%	15.8%	13.5%
CNOT Reduction	20.2%	35.5%	38.1%	43.2%	37.1%	32.1%

and CNOT count reduction. The results summary reaffirms the evaluation that the Population with Error and Cascade approaches perform the best, with an average improvement in TVD of 18.2% and 14.4%, respectively. Similar results are seen for JSD, with an average improvement of 15.8% and 13.5%, respectively. The Quest method achieves an average reduction in TVD and JSD of only 6.8%, giving the Population with Error method an advantage of 10% on both metrics. In terms of CNOT count reduction, the error aware methods are predictably among the better performing approaches, although the population-based approach does the best. The Cascade and Basic approaches are comparable, while the Quest suffers a drop of roughly 12% in comparison with the next closest method.

IV. Discussion

The extremely varied performance across most of the algorithms for most of the benchmarks raises a number of questions. The first concern is the relatively poor performance of the supposedly "enhanced" Quest approach (Basic), particularly when performance is below the original method. In these cases, the performance drop is caused by the improvement to the approximation limitation metric. The original algorithm's exploration of the search space is so significantly limited by the the faulty approximation metric that in most cases, particularly on simpler circuits, only a few circuits are returned. The improved metric does not suffer from this problem, but in several of these cases the additional circuits which are found are not of good quality. The Basic with Error configuration suffers a similar problem compared with Population with Error, as population with error is allowed to return the same circuit more than once, where Basic with Error is not. Thus, Basic with Error is occasionally forced to return poor quality circuits. The cause for the poor performance of the Cascade configuration on several circuit is likely due to the cascade metric breaking down and not providing significant benefits for circuits with few partitions. In these cases, similar results to Quest and the Basic method are expected and are indeed observed.

V. Conclusion

Full-circuit peephole optimization provides an interesting method for producing error resilient approximations of a given circuit which do not deviate too significantly from the exact output. However, limitations in the recombination step of existing methods must be addressed before these methods can be applied to larger quantum circuits. Notably, the recombination method proposed in Quest [3] has several shortcomings, including difficulty balancing correctness and complexity reduction, difficulty propagating approximation errors through circuits, sub-optimal differentiation metrics, and poor performance on circuits which have been mapped to restricted hardware. We address each of these problems by proposing changes to the recombination objective function,

with the best performing change seeing an 15% decrease in Total Variational Distance (TVD) and a 13.1% decrease in Jensen-Shannon Divergence (JSD) over the exact circuit. This corresponds to an 11.4% and 9.0% improvement over Quest in TVD and JSD, respectively.

Although the proposed methods provide good improvements over Quest, there is still much room for improvement. While the Population with Error and Cascade methods perform well, their performance is still quite poor on several circuits, and very inconsistent. We suspect that an error aware method with cascaded evaluation might perform well, but implementation of such a method is complicated by the fact that the Cascade method operates on circuit unitaries, while the error aware methods estimate error using the probability density matrices produced by each circuit. Further, the behavior of all methods regarding the differentiation metric is troubling, with most seeing little change in performance with the weight of the metric being reduced. Finally, we suspect that poor performance on the Adder 9 and Multiplier 10 benchmarks may be due to poor approximation quality, meaning that improvements in the approximate circuit generation may produce significant performance improvements.

References

[1] X.-C. Wu, M. G. Davis, F. T. Chong, and C. Iancu, *Qgo: Scalable quantum circuit optimization using automated synthesis*, 2020. DOI: 10.48550/ARXIV.2012.09835. [Online]. Available: https://arxiv.org/abs/2012.09835.

[2] M. Weiden, J. Kalloor, J. Kubiatowicz, E. Younis, and C. Iancu, "Wide quantum circuit optimization with topology aware synthesis," in *2022 IEEE/ACM Third International Workshop on Quantum Computing Software (QCS)*, IEEE, 2022, pp. 1–11.

[3] T. Patel, E. Younis, C. Iancu, W. de Jong, and D. Tiwari, "Quest: Systematically approximating quantum circuits for higher output fidelity," in *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2022, pp. 514–528.

[4] E. Jones, T. Oliphant, P. Peterson, *et al.*, *SciPy: Open source scientific tools for Python*, 2001–. [Online]. Available: http://www.scipy.org/.

[5] B. N. Laboratory, *Bqskit*, https://github.com/BQSKit/bqskit, 2022.