**Response to referee comments for: AN11713 Robustly decorrelating errors with mixed quantum gates**

We’d like to begin by thanking the referees for their detailed comments. We have used their comments and criticism to improve our manuscript and are now even more confident that it is suitable for publication in Physical Review A. Below we include the full report of each referee (in blue italics), interspersed with our response. Quotes from our revised manuscript are shown in green italics.

***Report of the First Referee -- AN11713/Polloreno***

*In this paper, the authors propose novel techniques to solve the problem of coherent errors in quantum computation. Their techniques, which are obtained by extending previous approaches, are based on various optimization methods. This topic is timely and the methods proposed here will be useful in the field of quantum information science. So, I recommend its publication in PRA.*

*Here I have a comment for readability. The presentation in Secs. IIIB3 and IVB1 with Fig. 6 may be difficult to understand, especially for readers not familiar with sparse modeling. I recommend the improvement of the parts.*

We thank the referee for their positive assessment of our work, as well as for alerting us to potentially unclear sections of the manuscript. In line with these comments, we have added the following text to Secs. IIIB3 and IVB1 and modified Fig. 6 to improve the clarity for readers not familiar with sparse optimization.

*Here, $\lambda \geq 0$ is a tunable parameter that can be used to control the degree of sparsity in the solution. If $\lambda$ is set to zero, the inner program reduces to Eq.~\eqref{eq:minimization} and the outer minimization is redundant. If $\lambda$ is very large, each inner minimization problem is minimized for an MQG with a single control, and the outer minimization simply selects the control which has the error generator of smallest norm. While this gives a maximally sparse solution, it doesn't allow for any cancellation of coherent errors.*

*In general, increasing $\lambda$ decreases the number of controls with large weights, and by removing controls with weights below a threshold probability we can increase the sparsity of the solution. Truncating in this way will introduce error in the Pauli-exactness of the solution. For larger $\lambda$ the required threshold probability decreases, and so the error introduced from truncation can be made smaller. However, larger $\lambda$ will also decrease the likelihood of the optimizer returning a generator-exact solution, and so in this way, tuning $\lambda$ allows us to trade between different sources of error when constructing sparse MQGs.*

**Report of the Second Referee -- AN11713/Polloreno**

The manuscript by Pollareno et al. presents a convex algorithm for compiling gates in quantum algorithms in a random fashion. Instead of applying a fixed gate pulse (which can have both coherent and incoherent errors), the authors pick the gates from ensembles with intentionally injected control errors. This procedure is meant to randomize coherent errors and mitigate any amplification of such errors in quantum circuits. The authors then performed a simple single-qubit RB experiment using their qubit device, where they show that the coherent errors of individual (intentionally misaligned) pulses are suppressed if the SQ gates are chosen from an ensemble of such misaligned gates. Some numerics are also done for two-qubit gates arguing for similar, beneficial outcomes.

Randomized compiling is technique known to the QC community for quite some time now and therefore there is nothing particularly novel about the technique itself. What appears to be the highlight of the paper is the convex algorithm that the authors used to choose random gates and the experimental demonstration. While these are good steps toward a nice direction, I feel the demonstrations are not sufficiently convincing, particularly in regard to the experimental part (which the authors themselves acknowledge as being "contrived"). Here are my main concerns:

While the referee is correct that our approach is superficially similar to randomized compiling, we wish to emphasize that our technique is intended as an *alternative approach* to solving the problem of coherent errors in quantum information processors. Our method is particularly well suited for experimental implementation, as the mixed unitary gates can be implemented at the level of individual logic operations using the FPGA control hardware. Contrast this against the randomized compiling approach, which requires wholesale recompilation of the circuit every time it is run. Importantly, though, we make no claim that this broad approach is novel – in fact, we frequently cite prior foundational work in this area. Instead, our manuscript extends and expands the scope of this prior work, formulating the problem as an explicit convex optimization, incorporating robustness to uncertainty, and, critically, providing the first experimental demonstration of this technique on quantum information processing hardware. We have added the following to the text to clarify this point:

MQGs are an alternative approach to solving this problem. Because they are implemented by randomly drawing from different gate realizations, they only require randomness at run-time. Moreover, when the construction of MQGs is posed as a numerical optimization problem, numerous useful modifications can be made to augment their performance.

We further thank the referee for their detailed assessment of our experimental demonstration. We address each of the referees comments below and, in doing so, we have substantially improved our manuscript.

1. For the experiment, it is completely trivial to me that randomly picking four faulty unitaries outperforms choosing just one of them. Here the only coherent errors the authors introduced are under and over-rotations to the Rabi angle. These over- and under-rotations would naturally build up and amplify if left by themselves. However, if chosen at random, they would cancel each other out and give better fidelity at the end of the sequence. Correct? I fail to see any surprise here and how the authors' lengthy algorithm even plays a role for this demonstration.

The referee is correct that Rabi oscillations would be significantly (and obviously) improved by switching to a mixed quantum gate implementation. In fact, we found this to be so obvious as to be uninteresting, so we chose not to test such circuits experimentally. What we did find surprising, however, was the improved performance of randomized benchmarking circuits. These circuits *already incorporate randomness* (as the referee points out below in comment 3)*.*  This randomness is specifically designed to – on average -- twirl coherent errors into stochastic errors. But the mixed quantum gate implementation both narrowed the variance over circuits and improved the average success probability. The former effect is a clear signature of reduced coherent noise, while the latter has a more complicated explanation: The improvement in success probability can be traced to the compilation of the Clifford operations, which occasionally contain sequences of repeated primitive gates. Coherent errors in these compilations can add up, increasing the average error rate of the compiled Clifford. When implemented using mixed quantum gates, the coherent addition of error is suppressed and the average compiled Clifford error rate is lower, leading to an improved success probability.

While the algorithm section of our manuscript is lengthy, it is central to our demonstration as it is used to choose the sampling weights for each of the erroneous implementations. If we had picked different weights, the effective channel would not be stochastic, and we would suffer a performance penalty in our demonstration.

2. It seems to me that authors can very easily go a step further: A SQ unitary is parametrized by three angles: Rabi angle, off-diagonal and diagonal phases. Why not introduce coherent errors to all three phases (or at least two, such as the Rabi angle and the off-diagonal phase)? To me this will be much less trivial.

This is an excellent suggestion, and we considered broadening the scope of the injected errors. However, we decided against this so that the presentation might be more streamlined. And while we agree that this might ultimately yield a more convincing demonstration, we no longer have the low-level access to Rigetti Computing hardware that would be required to implement it.

3. The whole point of RB is meant to randomize coherent errors and turn them into a depolarization channel. It is bizarre to me that the authors chose to demonstrate their randomized compiling through an experiment that is in itself random. Wouldn't a structured algorithm (QAOA or anything else that is not random) be a far better candidate for demonstrating the efficacy of randomized compiling?

As we discuss in our response to the referee’s first enumerated comment, randomized benchmarking circuits are sensitive to coherent errors. Demonstrations using metaheuristics such as QAOA, while currently *en vogue*, may instead confound the coherent error suppression we are attempting to illustrate. Full-stack QAOA requires a classical optimizer that continually tunes circuit parameters, so coherent errors can easily be suppressed by the optimizer. And, as we mentioned above, Rabi oscillation experiments just seemed *too* trivial to be interesting.

4. Why is the two-qubit experiment not done? It shouldn't be that difficult to do, if the authors believe that their method works (according to numerics). SQ demonstrations just too trivial to show anything given their simplicity.

We agree with the referee that a two-qubit gate demonstration would be significantly more impressive. Ultimately, we ourselves would like to see this method implemented on a two-qubit gate. However, the two-qubit gate available to us was dominated by large stochastic errors that severely curtailed the number of sequential gate applications before the state decohered. Injecting additional coherent errors would reduce this further. We felt that any convincing demonstration would be impossible to implement.

5. I'm still confused about how the algorithm works. Suppose we have gates that have systematic errors on some unitary angles. These errors are, of course, not known to me. Now, what do I do exactly? I start introducing a bunch of (additional and known) random errors to the gates, then compile the quantum circuit by picking gates from this random ensemble I constructed? If so, what the authors seemed to have showed is that the error of the randomly compiled circuit beats that of the circuit made from any one of the faulty gates in the ensemble (which I constructed myself), NOT the coherent error in the original gate. This seemed rather useless, as the endgame is to suppress the original, unknown coherent error, NOT the artificial errors I introduced. Am I missing something?

While the precise coherent errors experienced by a particular gate implementation might be unknown, it can certainly be learned fairly accurately using, for instance, a tomographic protocol. However, if our approach demanded such detailed knowledge it would be an operationally useless, though perhaps interesting, curiosity – you could simply calibrate away any errors you identify. Of course, it could be the case that the classical control hardware is sufficiently inflexible that some lingering coherent error will always remain (maybe you don't have enough samples in your DAC), or that the coherent errors come from unknown interactions that are out of your control. But an important enabling contribution of our work is the explicit inclusion of robustness into the optimization. This enables the mixed unitary gate to be protected against drift *or uncertainty in the precise nature of the coherent error.* This means that tomographic reconstruction of the error is unnecessary.

In our experiment we have introduced the miscalibrations intentionally, since the qubits were coherence limited. We have clarified this in the text as a summary of the protocol:

To construct an MQG, there needs to be several different realizations of the desired gate available. To see improved performance, these realizations need to be imperfect in different ways, from drift, miscalibration, or other sources of error. Because the qubit used in our demonstration was coherence-limited, and therefore did not suffer from miscalibration errors, we intentionally introduced coherent errors to showcase our technique.

By implementing four miscalibrated $X\_{\frac{\pi}{2}}$ rotations on a superconducting qubit and drawing from them with optimized weights, we see that coherent over- and under- rotations can be averaged away.

All in all, I cannot recommend the publication of the manuscript until the authors clarify my concerns above.

We believe that we have thoroughly addressed the referee’s concerns. We thank the referee for their observations, and we believe that responding to them has helped us to substantially improve our manuscript.