npj | quantum information

Article

Published in partnership with The University of New South Wales

https://doi.org/10.1038/s41534-024-00893-y

Gate-set evaluation metrics for closed-loop optimal control on nitrogen-vacancy center ensembles in diamond

Check for updates

Philipp J. Vetter[1,2,7], Thomas Reisser[3,4,7], Maximilian G. Hirsch[1,2,6], Tommaso Calarco[3,4,5], Felix Motzoi[3,4], Fedor Jelezko[1,2] & Matthias M. Müller[3]

A recurring challenge in quantum science and technology is the precise control of their underlying dynamics that lead to the desired quantum operations, often described by a set of quantum gates. These gates can be subject to application-specific errors, leading to a dependence of their controls on the chosen circuit, the quality measure and the gate-set itself. A natural solution would be to apply quantum optimal control in an application-oriented fashion. In turn, this requires the definition of a meaningful measure of the contextual gate-set performance. Therefore, we explore and compare the applicability of quantum process tomography, linear inversion gate-set tomography, randomized linear gate-set tomography, and randomized benchmarking as measures for closed-loop quantum optimal control experiments, using a macroscopic ensemble of nitrogen-vacancy centers in diamond as a test-bed. Our work demonstrates the relative trade-offs between those measures and how to significantly enhance the gate-set performance, leading to an improvement across all investigated methods.

Be it quantum computing, information, communication or metrology, precise control over a quantum system is a prerequisite for any successful application. To ensure robust results and the ability to run long quantum circuits, a great effort is spent by, e.g., IBM and Google on a periodic re-calibration of their quantum chips[1–6]. To this end a plethora of error correction and error suppression schemes are developed to achieve fault tolerant quantum computing[7–12]. Another example is quantum sensing, where the measured signals can be biased by quantum error correction[13] or distorted by faulty gates that are not accounted for[14].

In general, the precision of the system often suffers under, e.g., erroneous operations, wrongly populated states or limited knowledge of the system. Moreover, these errors must not necessarily accumulate linearly[15–17], leading to a dependence of the optimal controls for specific operations on the chosen circuit and the corresponding point of time within the selected circuit[2,3]. Thus, the quality of an operation has to be viewed within the context of its application, rather than looking at isolated individual gates; the focus should be shifted to their performance with respect to the chosen circuit and the entire gate-set. The ideal scenario would be to find a set of gates that performs universally well, regardless of the planned experiment. Since such a gate-set will heavily depend on the chosen experimental system, Quantum Optimal Control (QOC)[18–22] is well suited to tackle this complex problem.

Several optimization algorithms[23–31] are available in dedicated software packages such as our Quantum Optimal Control Suite (QuOCS)[32] that allows for closed-loop black-box optimization employing the dCRAB algorithm[22,27] via an interface to the lab software Qudi[33]. Such closed-loop optimizations based on measurements on the quantum system have been employed for gate optimization in superconducting qubits via randomized benchmarking[34,35], to optimize the preparation and phase transitions of Bose-Einstein condensates[36,37], enhance the macroscopic hyperpolarization in pentacene-doped naphthalene crystals[38] and for autonomous calibration of single-qubit gates as well as robust magnetometry of nitrogen-vacancy (N-*V*) centers in diamond[39,40]. For N-*V* centers in general, QOC has become a valuable tool for a wide range of challenges[14,21,41–49].

Room-temperature accessibility and control of their electronic spin state makes N-*V* centers ideal sensors for magnetic and electric fields[50–53]. Ensembles of N-*V* centers have particularly become a focus of attention, as these allow to drastically improve the sensitivity[54–56], create spatially resolved images of, e.g., local magnetic fields[57–59] and can even be used to create new phases of matter[60,61]. Due to their macroscopic size, many individual

[1]Institute for Quantum Optics, Ulm University, Ulm, Germany. [2]Center for Integrated Quantum Science and Technology (IQST), Ulm, Germany. [3]Peter Grünberg Institute – Quantum Control (PGI-8), Forschungszentrum Jülich GmbH, Jülich, Germany. [4]Institute for Theoretical Physics, University of Cologne, Cologne, Germany. [5]Dipartimento di Fisica e Astronomia, Università di Bologna, Bologna, Italy. [6]Present address: NVision Imaging Technologies GmbH, Ulm, Germany. [7]These authors contributed equally: Philipp J. Vetter, Thomas Reisser. e-mail: philipp.vetter@uni-ulm.de

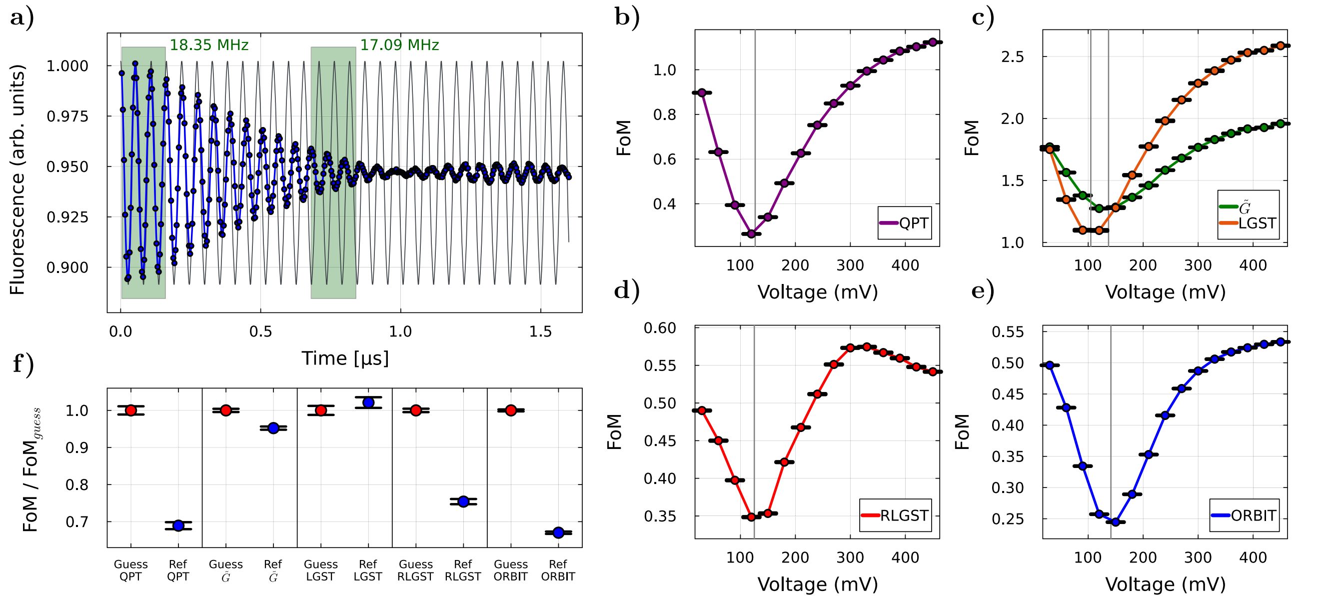

Fig. 1 | Ensemble dynamics and FoM validity measurements. a Rabi experiment with the N-V center ensemble, showing a strong beating due to the different Rabi oscillations of all individually contributing N-V centers. The green areas highlight the time dependence of the Rabi frequency and the black background oscillation visualizes the corresponding shift. **b** Linear sweep of the guess pulse amplitude for QPT, (**c**) LGST and $\tilde{G}$, (**d**) RLGST and (**e**) ORBIT. The ordinate shows the absolute value of the mean FoM from 20 measurements and the uncertainty. The minimum of the corresponding FoM is indicated by a gray line. **f** Comparison of the FoM definitions for the guess (red) and reference pulse (blue), normalized by FoM$_{\text{guess}}$. Each data point shows the mean value of 20 measurements with the corresponding standard deviation.

N-V centers contribute to the measurement, which ensures a very fast signal acquisition but can also result in strong detuning and large amplitude errors. These errors can negatively influence gate performances especially at different time-scales, as can be seen in Fig. 1a. The observed beating arises from the mixture of varying Rabi oscillation frequencies contributing to the measured signal. A macroscopic ensemble of N-V centers thus combines the effects of circuit- and length-dependent errors, large state preparation and measurement (SPAM) errors, a non-Markovian noise environment, and a high sensitivity to possible experimental drifts and distortions of the control pulses, making it an ideal test-bed to investigate the applicability of QOC for these kinds of inhomogeneous error mechanisms.

In order to find a universally well performing gate-set, a good measure is required, that reflects the quality of operations on the system in a holistic framework, being sensitive to the system dynamics on different time-scales and reliably accessible via measurements. To this end, we derive several measures of our gate-sets performance based on classical Quantum Process Tomography (QPT)[39,62–64], Linear-inversion Gate-Set Tomography (LGST)[65,66] and Randomized Linear Gate-Set Tomography (RLGST)[67] as well as Optimized Randomized Benchmarking for Immediate Tune-up (ORBIT)[35]. We experimentally asses their applicability for closed-loop optimizations in our test-bed and investigate how they reflect the system dynamics. We proceed by performing several optimization runs per method and cross-evaluate the performance of the optimized gate-set against all other methods as well as by Randomized Benchmarking (RB)[68–70]. Large improvements of the gate-set are observed in almost all optimization scenarios, significantly outperforming the provided guess and even outperforming the commonly used fastest possible rectangular shaped pulse. Moreover, we investigate how the chosen circuit length influences our obtained results for selected methods. Finally, we discuss the relative trade-offs of the individual methods applied on these types of systems and show how to best enhance the performance of the selected gate-set to find a universally well performing set.

Results

The experiments are performed with a CVD-grown diamond with a natural abundance (1.1%) of ^{13}C nuclei and a 10 µm thick layer of N-V centers. The layer contains an N-V concentration of roughly 1.5 ppm and is excited by a 532 nm laser with a beam waist of ≈38 µm. We apply a static magnetic field of 572 G to polarize the inherent nitrogen spin[71] and to create an effective qubit between the $|0\rangle$ and $|-1\rangle$ states. Gates are generated by microwave pulses sent through a straight microwave antenna made out of gold and placed on top of the diamond.

The figures of merit

We choose the gate-set

$$\begin{aligned}\mathcal{G} &= \{G_0, G_1, G_2, G_3, G_4, G_5, G_6\} \\ &= \{\mathbb{1}, \mathcal{X}_{\pi/2}, -\mathcal{X}_{\pi/2}, \mathcal{Y}_{\pi/2}, -\mathcal{Y}_{\pi/2}, \mathcal{X}_{\pi}, \mathcal{Y}_{\pi}\}\end{aligned} \tag{1}$$

for our experiments. Here, $\mathbb{1}$ is the identity and $\mathcal{X}$ and $\mathcal{Y}$ are the rotations around the x- and y- axis, where the rotation angle is given by the index. The gate-set is chosen such that we can generate the full Clifford group[72] needed for RB and ORBIT. We limit our optimizations to $G_2 = -\mathcal{X}_{\pi/2}$ to analyze how the change of an individual pulse affects the performance of our gate-set Eq. (1) and to ensure that the optimization converges quickly. All other gates are kept constant throughout all experiments, corresponding to rectangular pulses at maximum amplitude, whose pulse length is determined by a Rabi experiment. Except for QPT, all methods evaluate the performance of the entire gate-set, leading to the exact same measurements as if we would optimize all gates at once. Thus we do not expect a loss of generality of our results by limiting ourselves to one optimized gate. The initial state is set to $|\rho\rangle\rangle = |0\rangle\rangle$ and the POVM to $|E\rangle\rangle = |0\rangle\rangle$ for all of our experiments. Note that operators and states are given in Hilbert-Schmidt space to benefit from a simplified syntax (see supplementary information).

The performance of the gate-set is evaluated through measures based on QPT[39,62–64], LGST[65,66], RLGST[67] and ORBIT[35]. Additionally, we consider an adaption of LGST where we only take the targeted expectation values into account and label it $\tilde{G}$. The information extracted from these schemes is condensed down to a single real-valued number, called the Figure of Merit (FoM). The derivation and exact formula of the FoMs for each scheme can be found in the Methods "Definitions of the Figures of Merit". Note that the FoM is minimized in all experiments.

Different types of errors can be present in the system. Incoherent errors, as, e.g., bit-flip errors[73], are stochastic errors that do not preserve the purity of the state[74]. They often arise from the interaction with the environment or noise on the control fields, leading, e.g., to non-unitary processes. Incoherent noise effects scale linearly with the circuit size[75,76]. Coherent errors, on the other hand, arise from, e.g., deterministic interactions of the system with surrounding spins and constant electromagnetic fields or imperfections due to miscalibrations of the control field[77]. In contrast to incoherent errors, they scale quadratically with the circuit size[75,76]. Furthermore, the system can be subject to errors classified as Markovian and non-Markovian[78], which can be contained in both, coherent and incoherent effects[79]. Errors which exhibit temporal correlations and memory effects are referred to as non-Markovian errors, while Markovian errors are given by irreversible, memory-less errors, e.g., white noise.

Not all of our FoMs are susceptible to these error sources in the same way. QPT, $\tilde{G}$ and LGST, as employed here, will show a much stronger change of their corresponding FoM due to coherent errors in the form of over- and under-rotations than the other methods (see supplementary information). Due to their short circuit length, the correction of incoherent errors, as present in our system, like e.g., dephasing, will only lead to a slight change of their respective FoM. For RB on the other hand, the twirling condition[69] makes the method less sensitive to the full extend of coherent errors[74,80,81] and will thus weigh them differently. While ORBIT is based on RB, the method, and thus our FoM, directly measures the survival probability at a given circuit length. ORBIT does not perform a full-fledged RB experiment and therefore, a change of the coefficients that absorb SPAM errors or affect the final Clifford gate[70,82] can also lead to a change in our FoM. Despite the method, as employed in our studies, being mainly sensitive to incoherent errors, due to large circuit lengths, the FoM for ORBIT is still capable of detecting coherent errors. While RLGST is also based on randomized circuits, the method provides estimates of the gate-set, which is used to calculate our FoM. These estimates are only valid within the approximation of a linear-noise regime[67] and thus may suffer under large coherent errors. Due to the (potentially) significantly longer circuits, the method will be more sensitive to incoherent errors than the other methods with short circuit lengths. Like for LGST, incoherent errors can lead to a non-physical representation of the estimated gates. However, this must not be a problem for our application, as long as the FoM deteriorates for worse gates. In the case of RB and ORBIT, incoherent errors simply contribute to the observed exponential decay and they are also covered by the description of a process matrix in QPT.

Non-Markovian errors on the other hand will affect all FoMs. They can lead, for example, to a non-exponential decay in RB[80] and will negatively influence LGST[83]. Not only will coupling to surrounding nuclear spins lead to potentially non-Markovian errors during our measurements, but as we sum up the fluorescence signal of all N-V centers within our ensemble, specific sequences can exhibit strong temporal correlations, such as, e.g., seen by the beating of the Rabi oscillations in Fig. 1a. In our case, these errors arise from over- and under-rotations of the corresponding N-V centers attributed to the difference in driving field amplitude due to their varying distance to the microwave antenna, as well as their detunings due to the inhomogeneity of the external magnetic field. Thus, while some of our FoMs are by definition more sensitive to incoherent errors, due to their increased circuit length, these errors originate to a large degree from coherent dynamics, which if corrected, also lead to an improvement of the other FoMs. An effect that we more thoroughly analyze in "Optimizations with varying circuit length".

As QPT is unable to account for SPAM errors, all of the above mentioned errors can negatively influence the corresponding FoM even if they do not only act on the target gate.

During the closed-loop QOC experiments, QuOCS varies the S_x and S_y components of the microwave pulse amplitudes. To check the feasibility of our FoM definition, we sweep the pulse amplitude by varying the applied microwave voltage at a constant length $T = 30$ ns for a rectangular pulse, effectively creating a gate G_2 that either under- or over-rotates the spin instead of performing a $-\mathcal{X}_{\pi/2}$ operation. We select a circuit length of $L_R = 18$ for RLGST, which corresponds to a circuit length of $L_O = 10$ Cliffords for ORBIT. We average over 300 circuits to satisfy the twirling condition of RB and use the same number in RLGST and ORBIT for consistency, while ensuring that the number of circuits is significantly larger than the number of parameters to be estimated in RLGST[67]. The effect of this variation on the FoMs is shown in Fig. 1b–e. All curves show a distinct minimum from which the FoM increases upon deviation from the ideal voltage within the context of the individual technique. The obtained results again highlight the problem of circuit- and length-dependence of the optimal control parameters of our system, as we observe different minima, i.e., a different best amplitude, for all methods. This effect can be attributed to the previous discussion of coherent and incoherent errors and their influence on our FoMs.

Next, we investigate how a change in the pulse duration affects our FoMs. A short pulse with large amplitude is expected to outperform a long pulse with small amplitude because of its increased robustness against detuning errors as well as a smaller decoherence due to the shorter duration. In Fig. 1f the FoM of a 30 ns long rectangular pulse, which we from now on label as the "guess" pulse since it will be later handed to QuOCS as the guess at the start of our closed-loop QOC experiments, and the FoM of the shortest possible rectangular pulse of ≈14 ns, labeled as "reference" pulse, are shown. In both cases, the amplitude is determined through a Rabi experiment to match the selected pulse length. The shortest pulse is constrained by the maximal voltage that can be generated. Except for LGST, all methods show the expected behavior, that the short reference pulse achieves a significantly lower FoM than the guess pulse. While $\tilde{G}$ shows a smaller difference when compared with QPT, RLGST and ORBIT, the observed difference is well above the margin of error. LGST indicates that the reference pulse performs slightly worse than the guess pulse and it appears that the method cannot be used for our macroscopic ensemble as it does not capture the expected dynamics of our effective spin correctly. Therefore, we drop LGST as a measure for the cross-comparison of the optimized gate-sets between the different methods but still explore how it competes as a FoM for closed-loop optimizations.

Optimization gain

When comparing different analysis methods with each other, it is important to not only consider the absolute change in their FoM but to compare the relative change between two fixed reference points in order to correctly interpret the measurement results. Therefore, we introduce the so-called gain

$$\mathrm{Gain} = \frac{\mathrm{FoM}_{\mathrm{opt}} - \mathrm{FoM}_{\mathrm{guess}}}{\mathrm{FoM}_{\mathrm{ref}} - \mathrm{FoM}_{\mathrm{guess}}} \tag{2}$$

which displays the achieved improvement of the $\mathrm{FoM}_{\mathrm{opt}}$ in relation to the FoM of the reference and guess pulse. If the gain exceeds 1, it outperforms the rectangular reference pulse with maximum amplitude and if it falls below 0, the analyzed pulse performs worse than the provided guess pulse. The gain is a valuable quantity to ensure a fair comparison between the different analysis methods, regardless of their absolute value at any given point. If the absolute value of the FoM would change over several days due to thermal shifts or mechanical disturbances, the calculated gain remains constant.

Optimization workflow

The pulse shape of the microwave pulses

$$\Omega(t) = a_x(t)\cos(\omega t) + a_y(t)\sin(\omega t) \tag{3}$$

can be expressed by the time-dependent amplitudes $a_x(t)$ and $a_y(t)$, which represent the S_x and S_y (x- and y- spin operator) component of the corresponding gate. During our closed-loop QOC experiments we vary those time-dependent amplitudes according to the dCRAB algorithm[21,22,27,39] (see Methods "dCRAB Algorithm and Settings").

Figure 2 describes the workflow of such a closed-loop optimization. First, we select an analysis method, e.g., RLGST, which defines how the measurement sequence looks like and how the FoM is calculated. At the start of the optimization QuOCS[32] hands Qudi the initial guess pulse. The pulse is incorporated in the measurement sequence constructed by Qudi[33], our experimental control and data processing software. The full measurement sequence is uploaded via Qudi to the arbitrary waveform generator (Keysight M8195A) which generates the analog waveform and all required digital trigger signals. The analog waveform signal is sent through a 30 W amplifier (AR-30S1G6) at 80% gain and then through a straight gold microwave structure on the diamond's surface. The N-V centers are initialized and read out via a 532 nm green laser (Novanta Photonics gem 532) and their red fluorescence light is collected by a silicon photo-multiplier (Ketek PE3315-WB-TIA-SP). We use a digitizer (Spectrum Instrumentation M4i.4420-x8) to record the fluorescence signal which in turn is analyzed and evaluated by Qudi to obtain the N-V centers' spin state population. Once finished, Qudi calculates the FoM according to the selected analysis method, while treating the N-V ensemble as one effective qubit. The FoM is handed to QuOCS which in turn calculates a new pulse shape and the whole cycle repeats.

Following the closed-loop QOC experiment, the optimized pulse is benchmarked via all evaluation methods as well as a complete RB experiment.

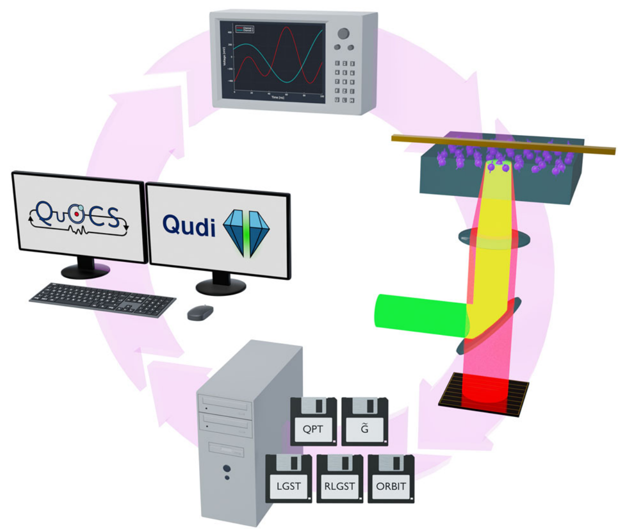

Fig. 2 | Workflow of the closed-loop QOC experiments. First, we select the analysis method, e.g., RLGST. Qudi then constructs the measurement sequence using the optimized pulse shape provided by QuOCS. The sequence is uploaded to the AWG to create the microwave signals. Finally, a green laser pulse is used to read out the N-V centers and the red fluorescence light is collected to calculate the corresponding FoM. The FoM is send to QuOCS which in return computes a new pulse shape and the cycle repeats.

Cross-comparison

Each optimization setting is repeated five times to investigate the stability and reproducibility of the optimization. The cross-evaluation of each experiment is shown in Fig. 3a displays an overview of all obtained gain values after optimization. Each column represents the method used for the FoM calculation during the closed-loop QOC experiment, each row represents the evaluation method and the height of the bars corresponds to the achieved gain, Eq. (2). For the evaluation with randomized benchmarking we use the extracted average error rate[69,70] instead of a FoM to calculate the corresponding gain. For a better comparison, we also show the average gain over all five optimizations for each method with the corresponding standard deviation in Fig. 3b. In Fig. 3c we show the gains of the pulse that performs best upon evaluation with its own optimization method for each of the five runs.

Optimization with QPT

For QPT, we observe a homogeneous gain progression across the whole column, i.e., for all evaluation methods. Evaluated with QPT, the achieved gain is on the same level as when evaluated by RLGST and RB. The first optimization run did not find a very good solution, which can happen because of the limited settings for the optimizer and can safely be marked as an outlier. Note, however, that this outlier leads to a large variance of the average values in Fig. 3b.

Optimization with $\tilde{G}$

Optimizing the pulse shape with $\tilde{G}$ leads to an outstanding improvement of the optimized pulse over the reference pulse when being evaluated with $\tilde{G}$ itself, resulting in an average gain of 1.34 and the highest achieved value of all

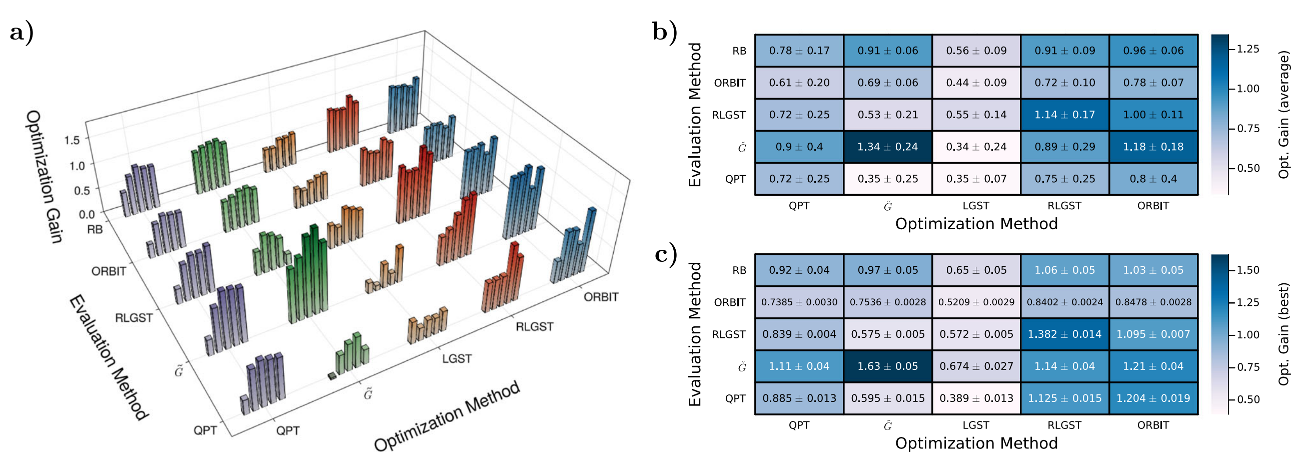

Fig. 3 | Cross-comparison of the closed-loop QOC experiments. a The x-axis shows the method selected to calculate the FoM during an optimization. We perform five optimization runs per method, represented by five bars, whose color denotes the optimization method. Their height corresponds to the achieved gain for the chosen evaluation method displayed in the y-axis and is is the average of 20 individual measurements. **b** Average gain of the five optimizations per method together with the standard deviation. The abscissa corresponds to the method selected for the FoM calculation during the optimization and the ordinate to the chosen evaluation method to calculate the gain. The color-scale indicated the achieved average gain. **c** The best performing pulse, when evaluated by its optimization method, is re-measured for all other evaluation methods. The error corresponds to the uncertainty of 20 measurements and the color-scale corresponds to the gain of the best performing pulse.

measurements of 1.63. However, the large improvement in the gain compared to the reference pulse is not retained for the other evaluation methods. We obtain low performances for QPT and RLGST but high gain values if evaluated with ORBIT and RB.

Despite their similarities, a large improvement in ORBIT must not translate to a large improvement in RLGST and vice versa, as they, e.g., weight coherent errors differently. The evaluation with RLGST, while still showing an improvement, displays the lowest overall achieved gain for all optimization methods. The same applies to the evaluation with QPT together with the LGST-optimized pulses.

A possible explanation is that $\tilde{G}$, ORBIT and RB do not provide estimates of the gates, while RLGST and QPT do. $\tilde{G}$ can find a compromise between the contributions of the expectation values to its FoM that better fits the pulse in the basis given by the rest of the gate-set while shifting the focus away from the action of G_2 itself. Eq. (5) for QPT is just a subset of the full set of expectation values measured in Eq. (7) for $\tilde{G}$. Depending on the measurement of all other expectation values, this specific set of expectations values of QPT is not necessarily maximized when the gate is optimized via $\tilde{G}$. Furthermore, while the expectation values of $\tilde{G}$ are gauge invariant, QPT is not. RLGST, as employed here in our system, does not have to be gauge invariant[67]. A slight adjustment of the gauge, e.g., shifts of the rotation axes, can thus worsen the FoM of QPT and RLGST while improving the one of $\tilde{G}$.

Optimization with LGST

Next, we analyze the optimizations with LGST via $\mathrm{FoM}_{\mathrm{LGST}}$ defined by Eq. (9). As expected, we observe the lowest gain for all evaluation methods across all optimization methods. Still, we always observe a positive gain, i.e., QuOCS is able to improve the gate-set's performance by varying the pulse shape of G_2. A major problem for optimization with LGST is that it is not able to differentiate between the reference and guess pulse as shown in Fig. 1f and might therefore also misinterpret good QOC pulses. Additionally, small variations in the pulse shape and measurement errors can lead to an enormous change of the FoM due to the calculation of the estimates via the Gram matrix in Eq. (8). The inversion of a matrix filled with measured, and therefore noisy, values as well as constantly changing entries due to the variations during the dCRAB search, is propagating and amplifying small changes through the whole derivation[84]. This results in a strongly fluctuating FoM during the optimization (see supplementary information). QuOCS can deal with such experimental downsides but it still influences the search for an optimal solution negatively.

Optimization with RLGST

Optimizing the pulse shape via RLGST shows a large average gain for all methods. The average value of 0.75 in Fig. 3b and the highest value with 1.13 in Fig. 3c when being evaluated with QPT even exceeds the QPT optimization itself. Despite operating at a different time-scale connected to the circuit length, RLGST is thus able to capture errors influencing the individual gate performance measured by QPT. The short circuit length of QPT suggests that those errors are coherent errors that accumulate with increasing circuit length, enabling a more precise optimization via RLGST. Moreover, we are able to find an optimized gate-set that reaches gains exceeding a value of 1 for almost all evaluation methods, as shown by Fig. 3c. The optimized pulse shape is thus capable of combating coherent and incoherent errors, which is further seen by the high gains when evaluated with ORBIT and $\tilde{G}$.

However, the optimization with RLGST also shows a large variance in the obtained gain when evaluated with QPT and $\tilde{G}$. While the method can capture the errors that dominate at short circuit lengths, not every single RLGST-optimized pulse will necessarily perform equally well or better than one being optimized with the corresponding FoM directly.

Optimization with ORBIT

For the optimization with ORBIT we observe a similar effect, showing a large improvement across all evaluation methods. If we take the average over all evaluation methods, ORBIT provides the best performance.

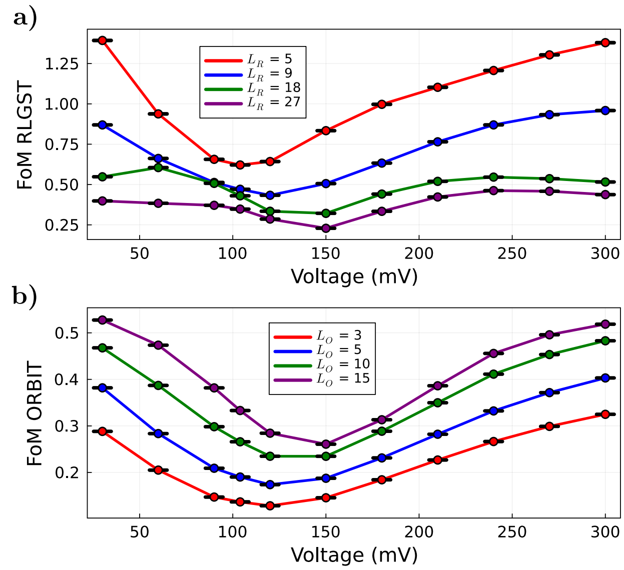

Fig. 4 | Dependence of the FoMs on the circuit length. Linear sweep of the guess pulse amplitude for different circuit lengths L for (**a**) RLGST and (**b**) ORBIT. The y-axis denotes the achieved FoM. Each data point represents the mean value of 20 measurements with the corresponding uncertainty and the chosen circuit length is visualized by the different colors.

For RLGST and $\tilde{G}$ we observe large improvements of the gain, reaching and even surpassing the performance of the reference pulse. The method also shows the highest average gain of all methods when evaluated with RB, which is only just below the reference pulse. Upon evaluation with QPT the optimized pulses on average outperform the ones achieved when being optimized via QPT itself.

Like the optimization with RLGST, the longer circuit lengths allows to better capture the main error contributions and optimize the pulse shape more precisely, leading to a robust gate-set at all evaluated circuit depths. Still, we observe a large variance between the different optimizations for the QPT evaluation. Again, the most plausible answer is that QPT is dominated by coherent errors in the form of over- or under-rotations, since incoherent errors only have a small influence on the FoM due to the short circuit length. As coherent errors are, to an extend, suppressed by twirling in RB[74,80,81] which partially translates to ORBIT, not every optimized pulse necessarily leads to a gate-set outperforming the reference pulse or the optimization with QPT itself. Additional measurements underlining this statement can be found in the supplementary information in sec. VII.

Circuit length dependence

The methods RLGST and ORBIT allow to easily probe how the system's dynamics at different time-scales affect the FoM by varying the selected circuit length. For this reason, we first examine how our defined FoMs for RLGST and ORBIT are affected by a linear change of the pulse amplitude for different L. The circuit length differs for the two methods: for RLGST it indicates the number of applied gates from the gate-set Eq. (1), while for ORBIT it stands for the number of applied Clifford gates. On average, one Clifford gate consists of ≈1.8 gates from our gate-set Eq. (1) and thus we choose the gate-string length of RLGST to match the average number of gates in ORBIT. Again we average over 300 random gate-strings per method. The measurement results are shown in Fig. 4a for RLGST and in Fig. 4b for ORBIT. We observe a shift of the minimum FoM, i.e., the best performing pulse amplitude, to larger amplitudes for increasing circuit length for both methods. This correlation can be explained by the strong beating observed in the Rabi experiment in Fig. 1a. Hopping from peak to peak of the Rabi oscillation is analogous to the application of a series of π-pulses on the system. If the π-pulse has a fixed pulse length, the amplitude of the pulse must be increased for subsequent applications to compensate for

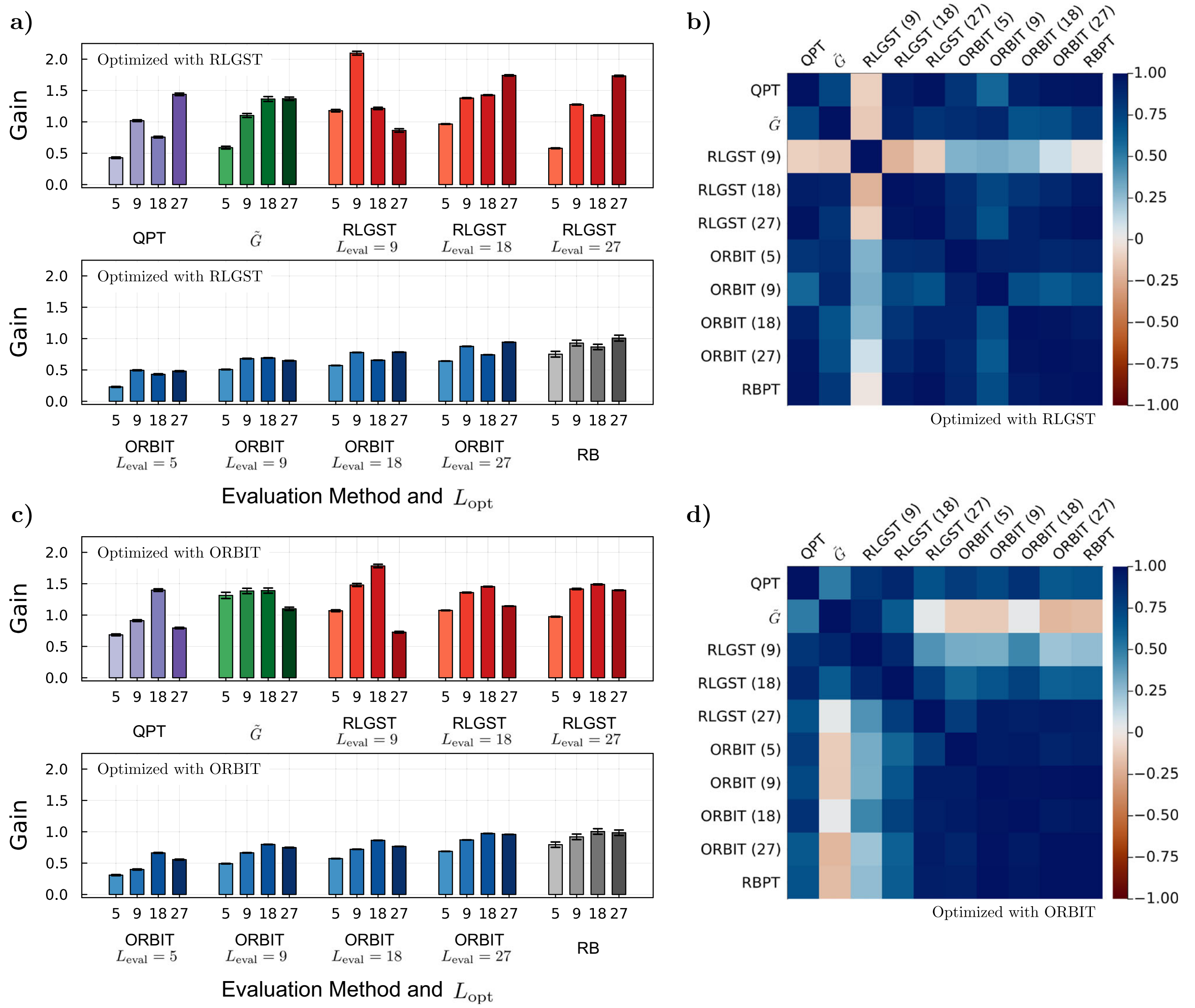

Fig. 5 | Influence of the circuit length on the optimization results. **a** Optimization runs with RLGST for four different circuit lengths $L = 5, 9, 18, 27$, represented by the four bars. The ordinate shows the gain achieved per optimization for the different evaluation methods displayed in the abscissa as an average of 20 measurements and the error bars denote the corresponding uncertainty. L_{eval} indicates the selected circuit length for the evaluation with RLGST and ORBIT. **b** Correlation matrix of the RLGST optimizations. The color-scale denotes the correlation between the different methods, i.e., if we observe the same dependence of the achieved gain on the circuit length. **c** Gain of the optimization runs for different circuit lengths with ORBIT. **d** Correlation matrix of the ORBIT optimizations. Again, the color-scale shows the correlation between the different methods.

the shift in the Rabi frequency caused by the beating. Thus, to find the best performing pulse for increasing number of repeated applications, a.k.a. L, the amplitude, in the form of a voltage in our case, needs to be adjusted upwards.

For ORBIT, the FoM gets worse (increases) with increasing circuit length as errors accumulate. This is expected as ORBIT evaluates the gate-set's performance at a fixed circuit length of the full RB curve, which exhibits an exponential decay with L. Therefore, increasing L samples the RB curve at a point of lower survival probability corresponding to a higher FoM per our definition. For RLGST we observe the exact opposite behavior, where an increased circuit length leads to a lower FoM. The FoM of RLGST in Eq. (14) in the Methods is based on the estimated error matrices for the initial state, POVM and the full gate-set, which are only accurate within the linear assumption[67]. Beyond this regime, the prediction error diverges[67] and the FoM decreases as the error matrices are under-estimated. Nonetheless, the argument of suitability of the FoM for our closed-loop QOC experiments holds due to the distinct minimum and the irrelevance of its absolute value as long as it guides the optimizer to a better solution. Importantly, both FoMs agree on the optimal amplitude for comparable L (minima of the curves with the same color code).

Optimizations with varying circuit length

Next, we investigate how a change of the circuit length affects the optimization result. A change in the circuit length can lead to a significant change in how different errors are weighted in the FoMs, e.g., the contribution of incoherent errors can be potentially increased through an increase of the circuit length. For example, L could be chosen such that it matches the intended circuit length of a desired experiment and thus might be tuned to give the best optimized gates for the given application. We therefore repeat the optimizations for RLGST and ORBIT once for different circuit lengths L. The results are shown in Fig. 5a for RLGST and Fig. 5c for ORBIT. We again choose the circuit length of RLGST such that it matches the average number of Clifford gates of ORBIT. For better readability we only display the average number of gates for ORBIT in Fig. 5c. The abscissa shows the chosen evaluation method, e.g., RLGST with a circuit length of $L_{\mathrm{eval}} = 18$, the ordinate shows the achieved gain. The four bars per evaluation method represent the individual optimizations for a chosen circuit length. We omit the evaluation with $L_{\mathrm{eval}} = 5$ for RLGST as the difference between the reference pulse and the guess pulse is within our measurement error and we are therefore not able to calculate the corresponding gain. Since we only perform one optimization per method and L_{opt}, we cannot exclude the possibility that some optimizations under- or over-performed due to the probabilistic nature of the optimization (caused by the measurement noise and the optimization algorithm). However, because of the small variance between different runs in Fig. 3, no significant change between repeated optimizations for the same L_{opt} is expected.

For both optimization methods, the overall achieved gain evaluated by ORBIT increases with increasing gate-string length L_{eval}. As the guess pulse is significantly longer than the reference pulse, the difference in their FoMs increases with the circuit length due to, inter alia, incoherent errors in the form of decoherence. This increases the distance between the FoMs of the reference and guess pulse. If the optimized pulse can compensate for this decoherence we observe an overall higher gain with increasing L_{eval} for the evaluation with ORBIT.

Figure 5b, d show the correlation matrices for the optimizations with different circuit lengths presented in Fig. 5a, c for RLGST and ORBIT, respectively. The correlation, described in the Methods "Correlation Matrix", looks at the behavior between the gains of different optimized pulses and compares it to other evaluation methods. E.g., if we observe that an increase of the circuit length leads to an increase of the gain when evaluated with ORBIT at $L = 27$, a high correlation with RB tells us that the method shows the same tendency.

When optimizing the gate-set via RLGST, we observe a high correlation between all evaluation methods except for the evaluation with RLGST at $L_{\mathrm{eval}} = 9$, as shown by the correlation matrix in Fig. 5b. An increase of the circuit length L_{opt} leads to an increase of the achieved gain. At large L_{opt}, small errors accumulate and cause a stronger change of the FoM, leading to a more precise and overall better optimization. This is also reflected by the absolute value of the gain, which for large L_{opt} almost always exceeds 1, when evaluated with RLGST, $\tilde{G}$ and QPT. For RLGST at $L_{\mathrm{eval}} = 9$ the gain even exceeds a value of 2, when optimized with the same circuit length. As the evaluation at this circuit length poorly correlates with all other evaluations, it can be assumed that the system exhibits certain dynamics exclusive to this time-scale and chosen measure.

The optimization with ORBIT at different circuit lengths shows again a high correlation between QPT, RB and the different ORBIT evaluations. The correlation matrix for ORBIT is shown in Fig. 5d. When evaluating the pulse with RLGST, we observe an increase of the correlation with ORBIT and RB for increasing L_{eval}. For $L_{\mathrm{eval}} = 27$, the observed dependence of the gain on the circuit length correlates very well with that of ORBIT. At large circuit lengths, the FoM's of RLGST and ORBIT are dominated by the same dynamics and thus converge to similar solutions, i.e., similar performing gate-sets. We observe again, that an increase of L_{opt} leads to a higher gain. Through the high correlation between the methods we can conclude that the optimization with ORBIT at $L_{\mathrm{opt}} = 27$ significantly under-performed and a higher gain is expected for repeated optimizations. Although showing high gains significantly exceeding 1, the evaluation with $\tilde{G}$ shows the smallest correlation with the other methods. Different circuit lengths during the optimization with ORBIT have thus no significant effect on the gain observed in $\tilde{G}$.

Evidently, it is favorable to perform optimizations with RLGST and ORBIT at higher L_{opt}. Apart from incoherent errors, a longer circuit length also allows for a higher precision when focusing on coherent errors prevalent at short time-scales. In our case, it does therefore provide no advantage to adapt L_{opt} to a desired experiment with, e.g., length L_{eval}, but it is simply better to choose the longest possible circuit length. Although this may seem surprising at first, it can be attributed to the ensemble nature of our system as we measure the collective fluorescence signal from all contributing individual N-V centers. The inhomogeneity of the microwave control field and external magnetic field heavily contribute to the observed dephasing (incoherent error) and apparent non-Markovianity in the Rabi measurement shown in Fig. 1a. Therefore, a pulse more robust to detuning and amplitude errors (coherent errors) can improve the performance for short and long sequences at the same time.

Discussion

Based on our detailed investigation, we can now discuss to what extent the individual methods are suitable for closed-loop optimizations with spin ensembles treated as one effective qubit or with qubits in general. We would like to emphasize that all methods have evidently been able to improve the performance of our gate-set.

Out of all investigated methods LGST achieves the lowest gains. Measurement errors dominate the provided estimates and therefore the calculated FoMs. More advance methods like long-sequence GST might help to overcome these shortcomings[83]. As a result, the optimizer sees no clear change for pulses that perform differently well, leading to an overall low optimization gain. If the FoM is calculated using solely the expectation values of LGST, i.e., $\tilde{G}$, significantly larger gains can be achieved. Since optimizations via $\tilde{G}$ maximize the overlap of the predefined expectation values to their target and allow to simply add additional gates to the gate-set, the method is particularly suitable if the gate-set is to be optimized for specific applications of the gates to arbitrary basis states. The method is heavily biased towards optimizing for its specific measurement sequence and the associated average circuit length.

QPT, unlike all the other methods, does not take SPAM errors into account. For larger SPAM errors, the correlation to other methods is expected to diminish as does its utility. Such is the case for our system, where an optimization based on QPT never beats the gate-set with the reference pulse. The optimization process is furthermore limited to one single gate at a time. While the optimization of an entire gate-set might be possible through iterative optimizations, the other methods offer a much better and clearer approach. Despite this, our FoM based on QPT is able to optimize the gate-set such that it shows improvements compared to the guess pulse for all investigated methods. Among all methods, QPT requires the fewest measurements, which can be beneficial when working with systems with low signal-to-noise ratio.

The overall best results are achieved by ORBIT. Not only does ORBIT show universally high gains across all methods, but the optimized gate-sets also outperform those of other optimization methods when evaluated with their own FoM, regardless of their time-scale. Although the method is mainly sensitive to incoherent errors, it also allows to correct those coherent errors that dominate the FoM for the other methods. As shown in the previous section, long circuit lengths are recommended for the optimization as errors accumulate, leading to larger changes in the defined FoM and therefore a more precise optimization. Too large L are not favorable, though, since information gets lost in noise and accumulated errors. When increasing the circuit length, we expect to find a sweet spot for the optimization where such dynamics are captured and weighed optimally. In our case we are limited by the fluence (see supplementary information) and therefore heating induced by long circuits.

One downside of ORBIT is that randomized benchmarking usually requires the use of Clifford gates which are generated by the over-full gate-set, Eq. (1). Additional gates that are to be optimized cannot simply be appended to the gate-set but need to be interleaved into the Clifford gates if possible.

This problem can largely be circumvented by the optimization via RLGST. Additional gates can simply be added to the gate-set. Optimizations via RLGST achieve similar high gains to ORBIT across all methods, i.e., are able to create a universally well-performing gate-set. Analogous to ORBIT, large circuit lengths are preferred when working with RLGST. At large circuit lengths, RLGST evaluates the system dynamics similar to ORBIT making the method ideal when working with a overfull gate-set as in our case. We show that the method can be used far beyond its intended linear (small error) working regime[67] to define a suitable FoM for closed-loop experiments. Moreover, while not directly investigated in our work, the method also allows to work with an incomplete gate-set[67], which can be of use for experimental systems with limited control. In addition, the method provides estimates of the gate-set and states which can be accessed during the optimization to obtain deeper insights into the system dynamics[85].

Nonetheless, as a large degree of our non-Markovian and incoherent errors originate from coherent dynamics due to the treatment of our micrometer-sized ensemble as one effective qubit, the recommendation for RLGST and ORBIT does not necessarily hold in other cases. For systems of a different nature such as ones measuring pure single qubits or physically different quantum platforms that have different error contributions, this might change. Short-circuit methods like QPT, $\tilde{G}$ and LGST could be selected over RLGST and ORBIT for other systems and shorter sequences in which the optimized gates are to be used. In general we expect, however, that RLGST and ORBIT with large L are suited well for applications that need longer circuit lengths. Further experiments that allow to strictly distinguish between coherent and incoherent errors[86] or even allow to, e.g., turn coherent errors into stochastic noise[74] would be of high interest to characterize the optimizations further or even optimize the gate-set with them.

In conclusion, through the definition of our FoMs based on RLGST and ORBIT we are able to create a gate-set that performs universally well across all investigated evaluation methods, regardless of their time-scale given by their circuit lengths and regardless of the different weighting of errors. We provide a detailed comparison of several gate and gate-set analysis methods and how these can be applied in a closed-loop QOC experiment, tested with a macroscopic ensemble of N-V centers. Since the occurring types of error are not exclusive to our system but exist in a wide variety of systems, our work can act as a useful handbook for experimentalists trying to overcome similar problems via QOC. The methods and the corresponding FoMs that work well for our system are expected to perform similar for other systems with comparable dynamics[59,87–90] and could also be applicable to systems with significantly smaller errors[91–93].

Methods

Definitions of the figures of merit

In addition to our gate-set Eq. (1), we define the SPAM-set to be

$$\begin{aligned}\mathcal{F} &= \{F_0, F_1, F_2, F_3\} \\ &= \{G_0, G_1, G_3, G_5\}\,.\end{aligned} \tag{4}$$

Starting with QPT, we measure the expectation values

$$p_{ij} = \langle\langle E|F_i G_2 F_j|\rho\rangle\rangle \tag{5}$$

for $F_{i,j} \in \mathcal{F}$ to evaluate the performance of G_2 along all axes and reconstruct the final states $|0\rangle$, $|-1\rangle$, $|+\rangle = 1/\sqrt{2}(|0\rangle + |-1\rangle)$ and $|-\rangle = 1/\sqrt{2}(|0\rangle + i|-1\rangle)$ after application of the gate. Finally, we calculate the process matrix χ according to[39,73], which defines the full linear map of our gate G_2. The FoM is then defined as the Frobenius norm of the distance between the measured and the target process matrix χ_{T} for the perfect gate:

$$\mathrm{FoM}_{\mathrm{QPT}} = \sqrt{\mathrm{tr}\Big(\left(\chi-\chi_{\mathrm{T}}\right)^{\dagger}\left(\chi-\chi_{\mathrm{T}}\right)\Big)}\,. \tag{6}$$

In contrast to QPT, quantum gate-set tomography[66,94] takes SPAM errors directly into account by measuring the expectation values of Eq. (5) for all $G_k \in \mathcal{G}$:

$$\begin{aligned}p_{ijk} &= \langle\langle E|F_i G_k F_j|\rho\rangle\rangle \\ &= (\tilde{G}_k)_{ij}.\end{aligned} \tag{7}$$

Following refs. 65,66, assuming a perfect idle gate $\mathbb{1}$ leads to the estimates

$$\begin{aligned}\hat{G}_k &= \tilde{G}_0^{-1}\tilde{G}_k, \\ |\hat{\rho}\rangle\rangle &= \tilde{G}_0^{-1}|\tilde{\rho}\rangle\rangle \\ |\hat{E}\rangle\rangle &= |\tilde{E}\rangle\rangle\,.\end{aligned} \tag{8}$$

using the inverse Gram matrix $\tilde{G}_0^{-1}$ to redistribute the errors to all other gates. As the expectation values (7) are gauge invariant, the estimates (8) need to be gauge-corrected. Following the derivations provided in refs. 65,66, we then obtain the gauge-transformed LGST estimates $\hat{G}_k^*$, $|\hat{\rho}^*\rangle\rangle$ and $|\hat{E}^*\rangle\rangle$. Usually, a maximum likelihood estimation (MLE) is used next to find the closest physically correct version of the obtained estimates, i.e., a $\hat{G}_k^*$ that corresponds to a CPTP map. However, such an MLE is computationally demanding, making it unsuitable for our closed-loop QOC experiments. We therefore use the estimates provided by LGST to evaluate the performance of our gates G_k through the difference to their target T_k:

$$\mathrm{FoM}_{\mathrm{LGST}} = \sqrt{\sum_{k=1}^{K+1}\mathrm{tr}\Big(\left(\hat{G}_k^*-T_k\right)^{\dagger}\left(\hat{G}_k^*-T_k\right)\Big)} \tag{9}$$

with $\hat{G}_{K+1}^* \equiv |\hat{\rho}^*\rangle\rangle\langle\langle\hat{E}^*|$. We sum up the squared norms for each gate difference similar to the sum of squared residuals in the least-squares method[95]. Additionally, we define another FoM

$$\mathrm{FoM}_{\tilde{G}} = \sqrt{\sum_{k=1}^{K}\mathrm{tr}\Big(\big(\tilde{G}_k-\tilde{T}_k\big)^{\dagger}\big(\tilde{G}_k-\tilde{T}_k\big)\Big)}\,, \tag{10}$$

based on the matrices $\tilde{G}_k$ of Eq. (7) filled with the measured expectation values and their difference to the corresponding target values $\tilde{T}_k$. While not providing an estimate, these expectation values are gauge invariant by definition.

Another method which takes SPAM errors into account is randomized benchmarking[68–70]. Randomized benchmarking uses Clifford gates, which are composed of gates from our gate-set Eq. (1)[72], to create multiple random circuits of length L, where the final gate flips the spin back to its initial state. The average recovered population, the so-called survival probability p_s, decays exponentially with the number of applied Clifford gates[69] and the average error per gate can be extracted from the decay parameter. ORBIT then allows to optimize the fidelity of the applied gates by increasing the survival probability at an arbitrary gate-string length L_{opt}[35]. Therefore, we define the corresponding FoM for our closed-loop QOC experiments with ORBIT by

$$\mathrm{FoM}_{\mathrm{ORBIT}} = 1 - p_s(L_{\mathrm{opt}})\,. \tag{11}$$

RLGST[67] combines the idea of gate-set tomography and randomized circuits to obtain an estimate of the gate-set with little computational overhead within a linear approximation. We start by measuring the expectation values

$$p_i = \langle\langle E|\mathcal{C}_i|\rho\rangle\rangle \tag{12}$$

for N randomly chosen circuits $\mathcal{C}_i$ of length L with $i = 1\ldots N$. Following the calculations provided in ref. 67, the expectation values of Eq. (12) are then used to obtain an estimate for the error matrices e_j, which describe how the measurement deviates from the expected target:

$$\begin{aligned}\hat{G}_k &= (\mathbb{1}+e_k)\cdot T_k, \\ |\hat{\rho}\rangle\rangle &= (\mathbb{1}+e_\rho)\cdot|\rho_T\rangle\rangle, \\ |\hat{E}\rangle\rangle &= (\mathbb{1}+e_E)\cdot|E_T\rangle\rangle\,.\end{aligned} \tag{13}$$

Since we only vary a single gate during our optimizations, we expect to be reasonably close to the actual gauge for the estimates and defining the FoM for RLGST equivalent to Eq. (9) we obtain

$$\mathrm{FoM_{RLGST}} = \sqrt{\sum_j \mathrm{tr}\left(e_j^{\dagger} e_j\right)}. \tag{14}$$

Correlation matrix

The correlation matrices are calculated according to ref. 96 via

$$M = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & 1 & r_{23} & \dots & r_{2p} \\ r_{31} & r_{32} & 1 & \dots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \dots & 1 \end{pmatrix} \tag{15}$$

with the correlation coefficient

$$r_{jk} = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{ik} - \bar{x}_k\right)^2}} \tag{16}$$

where the bar indicates the average over the indexed column of the data matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}. \tag{17}$$

The rows (n) of X are the four optimizations with varying L, i.e., the four pulses that are compared, and the columns (p) describe the different evaluation methods used.

dCRAB algorithm and settings

The dCRAB method expands updates to the amplitudes in a randomly selected sub-set of functions sampled from a "chopped" basis. We use the Fourier basis

$$a_k(t) = g(t) + \sum_{j=1}^{N_{\mathrm{SI}}}\sum_{i=1}^{N_b} c_{1,i}^{j}\sin\left(\nu_i^j t\right) + c_{2,i}^{j}\cos\left(\nu_i^j t\right), \tag{18}$$

where $g(t)$ is the provided guess. The frequencies ν_i^j for the update pulses are randomly selected from the interval $\left[0, \frac{2\pi \cdot n}{T}\right]$ where we choose $n = 4$ as the maximum allowed number of oscillations of the pulse envelope over the course of the pulse duration T. Several super-iterations (SI) with newly randomized basis vectors restart the search process by adding new search directions to avoid getting trapped in local minima. They are added to the best pulse shape of the previous SI. In our case we perform $N_{\mathrm{SI}} = 3$ super-iterations. We use the Nelder-Mead search method for the expansion parameters and select two basis vectors ($N_b = 2$) per pulse per super-iteration for the basis expansion. In this way, the search parameters for the optimization can be kept to a minimum, to ensure a quick convergence. If a potential improvement of the FoM does not exceed the measurement error, QuOCS will re-evaluate the evaluation step up to three times to ensure an accurate interpretation of the parameter landscape[32]. We determine the standard deviation by measuring the FoM for the guess pulse 100 times. A super-iteration stops, if the improvement of the FoM within the last 200 evaluation steps does not exceed the standard deviation. To account for experimental drifts in the FoM, the current best pulse is re-measured every 30 min.

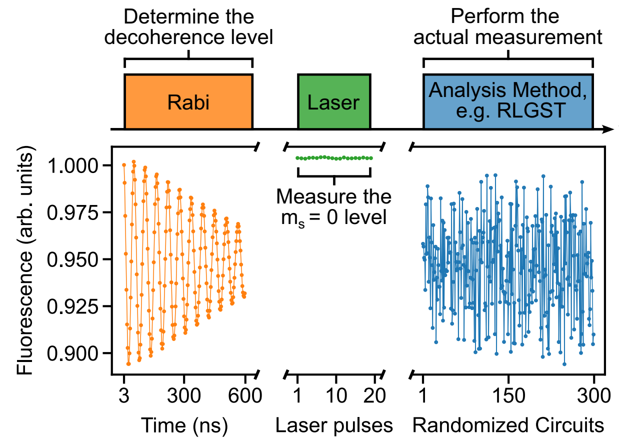

Fig. 6 | Measurement sequence. The measurement sequence for any experiment consists of three parts: a Rabi measurement to determine the decoherence level, a measurement of the $|0\rangle\rangle$ fluorescence level and the actual measurement sequence, e.g., RLGST, which is then converted to expectation values by the previously obtained fluorescence levels.

Data normalization and measurement sequence

For any performed experiment, the measurement sequence consists of three parts which are measured simultaneously, as depicted in Fig. 6. We start with a Rabi measurement at maximum amplitude for 600 ns to determine the decoherence level of our system and afterwards measure a series of 20 laser pulses to obtain the fluorescence level of the $|0\rangle\rangle$ state. Those levels are used to convert the fluorescence signal of the selected analysis method, to expectation values. Possible changes of the measurement contrast, through, e.g., laser power fluctuations, are thus taken into account by adjusting the measured expectation values such that 0 and 1 still correspond to the minimum and maximum achievable fluorescence.

Data availability

The data presented in this study is available from the corresponding authors on reasonable request.

Received: 21 March 2024; Accepted: 23 September 2024;
Published online: 02 October 2024

References

1. IBM. About calibration jobs. accessed: 19 January 2024. https://docs.quantum.ibm.com/admin/calibration-jobs.
2. Michielsen, K. et al. Benchmarking gate-based quantum computers. *Comput. Phys. Commun.* **220**, 44–55 (2017).
3. Papič, M., Auer, A. & de Vega, I. Fast estimation of physical error contributions of quantum gates. *arXiv preprint arXiv:2305.08916* (2023).
4. Tornow, C., Kanazawa, N., Shanks, W. E. & Egger, D. J. Minimum quantum run-time characterization and calibration via restless measurements with dynamic repetition rates. *Phys. Rev. Appl.* **17**, 064061 (2022).
5. Demirdjian, R., Gunlycke, D., Reynolds, C. A., Doyle, J. D. & Tafur, S. Variational quantum solutions to the advection–diffusion equation for applications in fluid dynamics. *Quantum Inform. Process.* **21**, 322 (2022).
6. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
7. Kelly, J. et al. State preservation by repetitive error detection in a superconducting quantum circuit. *Nature* **519**, 66–69 (2015).

8. Terhal, B. M. Quantum error correction for quantum memories. *Rev. Mod. Phys.* **87**, 307–346 (2015).
9. Chiaverini, J. et al. Realization of quantum error correction. *Nature* **432**, 602–605 (2004).
10. Cory, D. G. et al. Experimental quantum error correction. *Phys. Rev. Lett.* **81**, 2152–2155 (1998).
11. Cai, Z. et al. Quantum error mitigation. *Rev. Mod. Phys.* **95**, 045005 (2023).
12. Endo, S., Benjamin, S. C. & Li, Y. Practical quantum error mitigation for near-future applications. *Phys. Rev. X* **8**, 031027 (2018).
13. Rojkov, I., Layden, D., Cappellaro, P., Home, J. & Reiter, F. Bias in error-corrected quantum sensing. *Phys. Rev. Lett.* **128**, 140503 (2022).
14. Vetter, P. J. et al. Zero- and low-field sensing with nitrogen-vacancy centers. *Phys. Rev. Appl.* **17**, 044028 (2022).
15. Willsch, D., Nocon, M., Jin, F., De Raedt, H. & Michielsen, K. Gate-error analysis in simulations of quantum computers with transmon qubits. *Phys. Rev. A* **96**, 062302 (2017).
16. Lagemann, H. et al. Fragility of gate-error metrics in simulation models of flux-tunable transmon quantum computers. *Phys. Rev. A* **108**, 022604 (2023).
17. Marxer, F. et al. Implementing and characterizing high-fidelity two-qubit gates with long-distance transmon coupler. *Bull. Am. Phys. Soc.* **2023**, Z75–010 (2023).
18. d'Alessandro, D. *Introduction to quantum control and dynamics* (Chapman and hall/CRC, 2021).
19. Brif, C., Chakrabarti, R. & Rabitz, H. Control of quantum phenomena: past, present and future. *N. J. Phys.* **12**, 075008 (2010).
20. Glaser, S. J. et al. Training schrödinger's cat: quantum optimal control. *Eur. Phys. J. D.* **69**, 1–24 (2015).
21. Rembold, P. et al. Introduction to quantum optimal control for quantum sensing with nitrogen-vacancy centers in diamond. *AVS Quantum Sci.* **2**, 024701 (2020).
22. Müller, M. M., Said, R. S., Jelezko, F., Calarco, T. & Montangero, S. One decade of quantum optimal control in the chopped random basis. *Rep. Prog. Phys.* **85**, 076001 (2022).
23. Khaneja, N., Reiss, T., Kehlet, C., Schulte-Herbrüggen, T. & Glaser, S. J. Optimal control of coupled spin dynamics: design of nmr pulse sequences by gradient ascent algorithms. *J. Magn. Reson.* **172**, 296–305 (2005).
24. Machnes, S. et al. Comparing, optimizing, and benchmarking quantum-control algorithms in a unifying programming framework. *Phys. Rev. A* **84**, 022305 (2011).
25. Motzoi, F., Gambetta, J. M., Merkel, S. & Wilhelm, F. Optimal control methods for rapidly time-varying hamiltonians. *Phys. Rev. A* **84**, 022307 (2011).
26. Caneva, T., Calarco, T. & Montangero, S. Chopped random-basis quantum optimization. *Phys. Rev. A* **84**, 022326 (2011).
27. Rach, N., Müller, M. M., Calarco, T. & Montangero, S. Dressing the chopped-random-basis optimization: a bandwidth-limited access to the trap-free landscape. *Phys. Rev. A* **92**, 062343 (2015).
28. Fortunato, E. M. et al. Design of strongly modulating pulses to implement precise effective hamiltonians for quantum information processing. *J. Chem. Phys.* **116**, 7599–7606 (2002).
29. Ciaramella, G., Borzì, A., Dirr, G. & Wachsmuth, D. Newton methods for the optimal control of closed quantum spin systems. *SIAM J. Sci. Comput.* **37**, A319–A346 (2015).
30. Leung, N., Abdelhafez, M., Koch, J. & Schuster, D. Speedup for quantum optimal control from automatic differentiation based on graphics processing units. *Phys. Rev. A* **95**, 042318 (2017).
31. Preti, F., Calarco, T. & Motzoi, F. Continuous quantum gate sets and pulse-class meta-optimization. *PRX Quantum* **3**, 040311 (2022).
32. Rossignolo, M. et al. Quocs: the quantum optimal control suite. *Comput. Phys. Commun.* **291**, 108782 (2023).
33. Binder, J. M. et al. Qudi: A modular python suite for experiment control and data processing. *SoftwareX* **6**, 85–90 (2017).
34. Chow, J. M. et al. Optimized driving of superconducting artificial atoms for improved single-qubit gates. *Phys. Rev. A* **82**, 040305 (2010).
35. Kelly, J. et al. Optimal quantum control using randomized benchmarking. *Phys. Rev. Lett.* **112**, 240504 (2014).
36. Heck, R. et al. Remote optimization of an ultracold atoms experiment by experts and citizen scientists. *Proc. Natl Acad. Sci. USA* **115**, E11231–E11237 (2018).
37. Rosi, S. et al. Fast closed-loop optimal control of ultracold atoms in an optical lattice. *Phys. Rev. A* **88**, 021601 (2013).
38. Marshall, A. et al. Macroscopic hyperpolarization enhanced with quantum optimal control. *Phys. Rev. Res.* **4**, 043179 (2022).
39. Frank, F. et al. Autonomous calibration of single spin qubit operations. *npj Quantum Inf.* **3**, 1–5 (2017).
40. Oshnik, N. et al. Robust magnetometry with single nitrogen-vacancy centers via two-step optimization. *Phys. Rev. A* **106**, 013107 (2022).
41. Said, R. S. & Twamley, J. Robust control of entanglement in a nitrogen-vacancy center coupled to a ^{13}C nuclear spin in diamond. *Phys. Rev. A* **80**, 032303 (2009).
42. Scheuer, J. et al. Precise qubit control beyond the rotating wave approximation. *N. J. Phys.* **16**, 093022 (2014).
43. Dolde, F. et al. High-fidelity spin entanglement using optimal control. *Nat. Commun.* **5**, 3371 (2014).
44. Poggiali, F., Cappellaro, P. & Fabbri, N. Optimal control for one-qubit quantum sensing. *Phys. Rev. X* **8**, 021059 (2018).
45. Häberle, T., Schmid-Lorch, D., Karrai, K., Reinhard, F. & Wrachtrup, J. High-dynamic-range imaging of nanoscale magnetic fields using optimal control of a single qubit. *Phys. Rev. Lett.* **111**, 170801 (2013).
46. Nöbauer, T. et al. Smooth optimal quantum control for robust solid-state spin magnetometry. *Phys. Rev. Lett.* **115**, 190801 (2015).
47. Ziem, F., Garsi, M., Fedder, H. & Wrachtrup, J. Quantitative nanoscale mri with a wide field of view. *Sci. Rep.* **9**, 1–9 (2019).
48. Konzelmann, P. et al. Robust and efficient quantum optimal control of spin probes in a complex (biological) environment. towards sensing of fast temperature fluctuations. *N. J. Phys.* **20**, 123013 (2018).
49. Müller, M. M., Gherardini, S. & Caruso, F. Noise-robust quantum sensing via optimal multi-probe spectroscopy. *Sci. Rep.* **8**, 14278 (2018).
50. Balasubramanian, G. et al. Nanoscale imaging magnetometry with diamond spins under ambient conditions. *Nature* **455**, 648–651 (2008).
51. Mamin, H. et al. Nanoscale nuclear magnetic resonance with a nitrogen-vacancy spin sensor. *Science* **339**, 557–560 (2013).
52. Schmitt, S. et al. Submillihertz magnetic spectroscopy performed with a nanoscale quantum sensor. *Science* **356**, 832–837 (2017).
53. Dolde, F. et al. Electric-field sensing using single diamond spins. *Nat. Phys.* **7**, 459–463 (2011).
54. Glenn, D. R. et al. High-resolution magnetic resonance spectroscopy using a solid-state spin sensor. *Nature* **555**, 351–354 (2018).
55. Barry, J. F. et al. Sensitivity optimization for nv-diamond magnetometry. *Rev. Mod. Phys.* **92**, 015004 (2020).
56. Balasubramanian, P. et al. dc magnetometry with engineered nitrogen-vacancy spin ensembles in diamond. *Nano Lett.* **19**, 6681–6686 (2019).
57. Horsley, A. et al. Microwave device characterization using a widefield diamond microscope. *Phys. Rev. Appl.* **10**, 044039 (2018).
58. Mizuno, K. et al. Wide-field diamond magnetometry with millihertz frequency resolution and nanotesla sensitivity. *Aip Adv.* **8**, 125316 (2018).
59. Scholten, S. et al. Widefield quantum microscopy with nitrogen-vacancy centers in diamond: strengths, limitations, and prospects. *J. Appl. Phys.* **130**, 150902 (2021).
60. Choi, S. et al. Observation of discrete time-crystalline order in a disordered dipolar many-body system. *Nature* **543**, 221–225 (2017).
61. Choi, J. et al. Probing quantum thermalization of a disordered dipolar spin ensemble with discrete time-crystalline order. *Phys. Rev. Lett.* **122**, 043603 (2019).
62. Poyatos, J. F., Cirac, J. I. & Zoller, P. Complete characterization of a quantum process: The two-bit quantum gate. *Phys. Rev. Lett.* **78**, 390–393 (1997).

63. Chuang, I. L. & Nielsen, M. A. Prescription for experimental determination of the dynamics of a quantum black box. *J. Mod. Opt.* **44**, 2455–2467 (1997).
64. Leibfried, D. et al. Experimental determination of the motional quantum state of a trapped atom. *Phys. Rev. Lett.* **77**, 4281–4285 (1996).
65. Greenbaum, D. Introduction to quantum gate set tomography. *arXiv preprint arXiv:1509.02921* (2015).
66. Blume-Kohout, R. et al. Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit. *arXiv preprint arXiv:1310.4492* (2013).
67. Gu, Y., Mishra, R., Englert, B.-G. & Ng, H. K. Randomized linear gate-set tomography. *PRX Quantum* **2**, 030328 (2021).
68. Knill, E. et al. Randomized benchmarking of quantum gates. *Phys. Rev. A* **77**, 012307 (2008).
69. Magesan, E., Gambetta, J. M. & Emerson, J. Scalable and robust randomized benchmarking of quantum processes. *Phys. Rev. Lett.* **106**, 180504 (2011).
70. Magesan, E., Gambetta, J. M. & Emerson, J. Characterizing quantum gates via randomized benchmarking. *Phys. Rev. A* **85**, 042311 (2012).
71. Jacques, V. et al. Dynamic polarization of single nuclear spins by optical pumping of nitrogen-vacancy color centers in diamond at room temperature. *Phys. Rev. Lett.* **102**, 057403 (2009).
72. Epstein, J. M., Cross, A. W., Magesan, E. & Gambetta, J. M. Investigating the limits of randomized benchmarking protocols. *Phys. Rev. A* **89**, 062321 (2014).
73. Nielsen, M. A. & Chuang, I. *Quantum computation and quantum information* (Cambridge University Press, 2002).
74. Hashim, A. et al. Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor. *Phys. Rev. X* **11**, 041039 (2021).
75. Wallman, J. J. & Emerson, J. Noise tailoring for scalable quantum computation via randomized compiling. *Phys. Rev. A* **94**, 052325 (2016).
76. Iverson, J. K. & Preskill, J. Coherence in logical quantum channels. *N. J. Phys.* **22**, 073066 (2020).
77. Kim, J.-S. et al. Hardware-efficient random circuits to classify noise in a multiqubit system. *Phys. Rev. A* **104**, 022609 (2021).
78. Haase, J. F. et al. Controllable non-markovianity for a spin qubit in diamond. *Phys. Rev. Lett.* **121**, 060401 (2018).
79. Penshin, P. et al. Realization of robust quantum noise characterization in the presence of coherent errors. *AVS Quantum Science 6.2* (2024).
80. Proctor, T., Rudinger, K., Young, K., Sarovar, M. & Blume-Kohout, R. What randomized benchmarking actually measures. *Phys. Rev. Lett.* **119**, 130502 (2017).
81. Boone, K., Carignan-Dugas, A., Wallman, J. J. & Emerson, J. Randomized benchmarking under different gate sets. *Phys. Rev. A* **99**, 032329 (2019).
82. Magesan, E. et al. Efficient measurement of quantum gate error by interleaved randomized benchmarking. *Phys. Rev. Lett.* **109**, 080505 (2012).
83. Nielsen, E. et al. Gate set tomography. *Quantum* **5**, 557 (2021).
84. Lefebvre, M., Keeler, R., Sobie, R. & White, J. Propagation of errors for matrix inversion. *Nucl. Instrum. Methods Phys. Res. Sect. A: Accelerators, Spectrometers, Detect. Associated Equip.* **451**, 520–528 (2000).
85. Wittler, N. et al. Integrated tool set for control, calibration, and characterization of quantum devices applied to superconducting qubits. *Phys. Rev. Appl.* **15**, 034080 (2021).
86. Feng, G. et al. Estimating the coherence of noise in quantum control of a solid-state qubit. *Phys. Rev. Lett.* **117**, 260501 (2016).
87. Gottscholl, A. et al. Spin defects in hbn as promising temperature, pressure and magnetic field quantum sensors. *Nat. Commun.* **12**, 4480 (2021).
88. Gong, R. et al. Coherent dynamics of strongly interacting electronic spin defects in hexagonal boron nitride. *Nat. Commun.* **14**, 3299 (2023).
89. Castelletto, S. & Boretti, A. Silicon carbide color centers for quantum applications. *J. Phys. Photon.* **2**, 022001 (2020).
90. Joas, T. et al. High-fidelity electron spin gates in a scalable diamond quantum register. *arXiv preprint arXiv:2406.04199* (2024).
91. Bartling, H. P. et al. Universal high-fidelity quantum gates for spin-qubits in diamond *arXiv preprint arXiv:2403.10633* (2024).
92. Werninghaus, M. et al. Leakage reduction in fast superconducting qubit gates via optimal control. *npj Quantum Inf.* **7**, 14 (2021).
93. Dehollain, J. P. et al. Optimization of a solid-state electron spin qubit using gate set tomography. *N. J. Phys.* **18**, 103018 (2016).
94. Merkel, S. T. et al. Self-consistent quantum process tomography. *Phys. Rev. A* **87**, 062119 (2013).
95. Van Huffel, S. & Vandewalle, J.*The Total Least Squares Problem* (Society for Industrial and Applied Mathematics, 1991). https://epubs.siam.org/doi/abs/10.1137/1.9781611971002.
96. Helwig, N. E. Data, covariance, and correlation matrix. *University of Minnesota (Twin Cities)* (2017).

Acknowledgements
We thank Ressa Said, Timo Joas, Marco Rossignolo and Maximilian Kraus for helpful discussions. This work was funded by the German Federal Ministry of Research (BMBF) by future cluster QSENS (No. 03ZK110AB) and projects DE-Brill (No. 13N16207), SPINNING (No. 13N16210 and No. 13N16215), DIAQNOS (No. 13N16463), quNV2.0 (No. 13N16707), QR.X (No. 16KISQ006) and Quamapolis (No. 13N15375), DLR via project QUASIMODO (No. 50WM2170), Deutsche Forschungsgemeinschaft (DFG) via Projects No. 386028944, No. 445243414, and No. 387073854 and Excellence Cluster POLiS (UP 33/1 Projekt-ID: 422053626), Excellence Cluster Matter and Light for Quantum Computing (ML4Q) EXC 2004/1 - 390534, and Helmholtz Validation Fund project "Qruise" (HVF-00096). This project has also received funding from the European Union's HORIZON Europe program via projects QuMicro (No. 101046911), SPINUS (No. 101135699), C-QuENS (No. 101135359), QCIRCLE (No. 101059999), HORIZON-CL4-2021-DIGITALE-MERGING-02-10 QCFD (No. 101080085) and FLORIN (No. 101086142), European Research Council (ERC) via Synergy grant HyperQ (No. 856432) and Carl-Zeiss-Stiftung via the Center of Integrated Quantum Science and technology (IQST) and project Utrasens-Vir (No. P2022-06-007).

Author contributions
P.J.V. and T.R. contributed equally to the project. P.J.V., T.R., M.M.M., and F.M. conceived the experiments. P.J.V., T.R., and M.G.H. implemented the closed-loop optimizations. P.J.V. performed the measurement. P.J.V. and T.R. derived the figures of merit, analyzed the data and wrote the manuscript. M.M.M., F.M., T.C., and F.J. supervised the project. All authors read and commented on the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41534-024-00893-y.

Correspondence and requests for materials should be addressed to Philipp J. Vetter.