Tutorial

Practical Introduction to Benchmarking and Characterization of Quantum Computers

Akel Hashim,[1,2,*] Long B. Nguyen,[1,2] Noah Goss,[1,2] Brian Marinelli,[1,2] Ravi K. Naik,[1,2] Trevor Chistolini,[1] Jordan Hines,[1,3] J.P. Marceaux,[4,5] Yosep Kim,[6] Pranav Gokhale,[7] Teague Tomesh,[7] Senrui Chen,[8] Liang Jiang,[8] Samuele Ferracin,[9,†] Kenneth Rudinger,[3] Timothy Proctor,[4] Kevin C. Young,[4] Irfan Siddiqi,[1,2,10] and Robin Blume-Kohout[3]

[1] *Department of Physics, University of California at Berkeley, Berkeley, California 94720, USA*

[2] *Applied Math and Computational Research Division, Lawrence Berkeley National Lab, Berkeley, California 94720, USA*

[3] *Quantum Performance Laboratory, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA*

[4] *Quantum Performance Laboratory, Sandia National Laboratories, Livermore, California 94550, USA*

[5] *Graduate Group in Applied Science and Technology, University of California at Berkeley, Berkeley, California 94720, USA*

[6] *Department of Physics, Korea University, Seoul 02841, Korea*

[7] *Infleqtion, Chicago, Illinois 60604, USA*

[8] *Pritzker School of Molecular Engineering, University of Chicago, Illinois 60637, USA*

[9] *Keysight Technologies Canada, Kanata, ON K2K 2W5, Canada*

[10] *Materials Sciences Division, Lawrence Berkeley National Lab, Berkeley, California 94720, USA*

(Received 22 August 2024; published 15 August 2025)

Rapid progress in quantum technology has transformed quantum computing and quantum information science from theoretical possibilities into tangible engineering challenges. Breakthroughs in quantum algorithms, quantum simulations, and quantum error correction are bringing useful quantum computation closer to fruition. These remarkable achievements have been facilitated by advances in quantum characterization, verification, and validation (QCVV). QCVV methods and protocols enable scientists and engineers to scrutinize, understand, and enhance the performance of quantum information-processing devices. In this tutorial, we review the fundamental principles underpinning QCVV, and introduce a diverse array of QCVV tools used by quantum researchers. We define and explain QCVV's core models and concepts—quantum states, measurements, and processes—and illustrate how these building blocks are leveraged to examine a target system or operation. We survey and introduce protocols ranging from simple qubit characterization to advanced benchmarking methods. Along the way, we provide illustrated examples and detailed descriptions of the protocols, highlight the advantages and disadvantages of each, and discuss their potential scalability to future large-scale quantum computers. This tutorial serves as a guidebook for researchers unfamiliar with the benchmarking and characterization of quantum computers, and also as a detailed reference for experienced practitioners.

DOI: 10.1103/PRXQuantum.6.030202

CONTENTS

*Contact author: ahashim@berkeley.edu
†Now at IBM, Markham, ON, Canada.

List of Acronyms

ACES averaged circuit eigenvalue sampling
AGF average gate fidelity
AGI average gate infidelity
AGSI average gate-set infidelity
AWG arbitrary waveform generator
BiRB binary randomized benchmarking
BME Bayesian mean estimate
CB cycle benchmarking
CER cycle error reconstruction
CP complete positivity
CPTP completely positive trace-preserving
CRB Clifford-group randomized benchmarking
CW continuous-wave
DD dynamical decoupling
DFE direct fidelity estimation
DRB direct randomized benchmarking
EPC error per Clifford
EPLG error per layered gate
GHZ Greenberger-Horne-Zeilinger
GST gate set tomography
IRB interleaved randomized benchmarking
KL Kullback-Leibler
KNR *k*-body noise reconstruction
LGST linear gate set tomography
LRB leakage randomized benchmarking
LSGST long sequence gate set tomography
MAP maximum a posteriori
MCFE mirror circuit fidelity estimation
MCM mid-circuit measurement
MLE maximum likelihood estimation
MRB mirror randomized benchmarking
PB purity benchmarking
PLS projected least squares
POVM positive operator-valued measure
PTM Pauli transfer matrix
PVM projection-valued measure
QAOA quantum approximate optimization algorithm
QCVV quantum characterization, verification, and validation
QFT quantum Fourier transform
QI quantum instrument
QMT quantum measurement tomography
QND quantum nondemolition
QPT quantum process tomography
QST quantum state tomography
QV quantum volume
RB randomized benchmarking
RMS root mean square
RPE robust phase estimation
RWA rotating wave approximation
SPB speckle purity benchmarking
sRB simultaneous randomized benchmarking
SPAM state preparation and measurement
TP trace preservation
TVD total variation distance
XEB cross-entropy benchmarking
XRB eXtended randomized benchmarking

I. INTRODUCTION

Quantum computation has grown from a mere theoretical proposition [1] to a tangible reality, heralding a new era of science in quantum applications across diverse domains, with recent breakthroughs in quantum measurement and control [2–5], fundamental science [6–11], computer science [12–16], quantum chemistry [17], materials science [18], and many others. If quantum computers of sufficient size and precision can be built, they promise to deliver computational advantages in a diverse range of applications [1,19–26]. Much work remains to be done before these promises become a reality [27], but the progress of quantum processors over the past three decades suggests that success will eventually be achievable. That progress—both past, and future—is enabled and facilitated by the growing toolbox of *quantum characterization, verification, and validation (QCVV)*.

QCVV means the characterization and benchmarking of quantum computers and their constituent components. It encompasses a large set of methods, protocols, and concepts that have been developed over the past 30 years. These techniques probe the *in situ* behavior of qubits, quantum logic operations, and integrated quantum processors. Having done so, they report detailed predictive models of a device's behavior (*characterization*), simple figures of merit (*benchmarking*), or hybrids of the two. The majority of the QCVV literature focuses on gate-based quantum computers (rather than analog simulators, quantum annealers, or other paradigms), and this tutorial will too. QCVV of gate-based quantum computers is primarily concerned with characterizing and/or benchmarking the quantum states, quantum gates, and quantum measurements that are implemented by (multi)qubit devices. It is possible and useful to distinguish characterization protocols from benchmarking protocols. But, in practice these two disciplines are complementary and sometimes overlap (e.g., randomized benchmarking techniques can be deployed for either purpose), and they rely and build upon the same foundational concepts.

Put simply, the goal of QCVV is to learn about as-built quantum computing devices. This usually means using data to estimate properties of mathematical models for those devices, with the goal of predicting (either qualitatively or quantitatively) their future behavior. We conceptualize QCVV methods as tools in a large toolbox. Many of those tools are protocols that can be deployed by an experimentalist or engineer to obtain specific information about the behavior of a quantum computational device (e.g., a qubit, logic operation, or integrated processor) [28]. In this tutorial, we introduce the most common black-box models used to describe and predict the behavior of quantum computers. Using those models, we introduce the most commonly encountered failure modes of qubits and quantum logic gates, explain how they affect quantum computations, and survey the most common metrics used to quantify their impact. Then, in the bulk of the tutorial, we survey the most common QCVV methods and protocols used to learn models for quantum computer performance. In so doing, we discuss the advantages and disadvantages of each method, the trade-offs between them, their scalability, and their relative utility in predicting the behavior of quantum computations.

A. Uses of QCVV

Most QCVV protocols fall into one of four categories:

(1) physical device characterization,
(2) tomographic characterization,
(3) randomized benchmarks, and
(4) holistic (application-centric) benchmarking.

Qubits and quantum computers are (as of 2025) still mostly physics experiments. When a qubit or multiqubit device is fabricated, it cannot be treated or operated as a quantum computer until its physical properties—e.g., the resonant frequencies and coherence times of qubits, and the nature of couplings between them—have been determined, calibrated, and optimized. This is the domain of physical device characterization (see Sec. VI).

Once a quantum computational device has been calibrated, it becomes possible to treat (and model) it as a quantum computer rather than a physics experiment. Tomographic characterization is now possible. Tomographic QCVV protocols (see Sec. VII) aim to measure and reconstruct (or *estimate*) the state of one or more qubits, or the operation (e.g., logic gate or measurement) acting on them. For example, quantum state tomography (Sec. VII A) estimates the density matrix describing an initialization operation, while quantum process tomography (Sec. VII B) estimates the superoperator describing a reversible logic gate. Tomography-based methods are widely used to characterize individual components, but are generally not scalable to large quantum systems.

Randomized benchmarks (see Sec. VIII) are intended to build relatively *qualitative* assessments of quantum device performance. Randomized benchmarks probe the performance of an entire set of quantum logic gates and summarize it with $\mathcal{O}(1)$ numbers, without attempting to characterize or model each gate in detail. They report those gates' average performance over many possible input states and many possible contexts, thus providing the end user with some intuition (but few guarantees) about how well the gate will perform in different circuits. Randomized benchmarks are often used to probe just one or two qubits, but some are scalable and can be used to assess the overall performance of an entire processor. In all cases, they provide significantly less detailed and predictive information than tomographic protocols.

Holistic benchmarks (see Sec. XI) are intended to measure the performance of a quantum computer on "relevant" tasks. Like scalable randomized benchmarks, holistic benchmarks ignore the underlying details of individual qubits and gates. Some holistic benchmarks are designed to capture the performance of a quantum computer in a single number, while others seek to predict how well a quantum computer would perform at a range of different circuit depths and widths. Most holistic benchmarks are designed (unlike randomized benchmarks) to measure the performance of a specific application or class of algorithms.

All of these methods are useful tools in the QCVV toolbox. Our goal in this tutorial—in addition to teaching the foundational concepts and methods that underlie *all* QCVV protocols—is to enable readers to decide which QCVV method(s) to use. They are very different tools, and the best tool for a given job depends on the user's goals and needs. Each class of protocols (1) makes different assumptions, (2) seeks to gain a different amount or kind of information about the quantum device being probed, (3) is more or less scalable to large devices, and (4) is more or less amenable to rigorous certification [29,30]. In this tutorial, we aim to teach readers about these trade-offs, and to enable scientists and engineers to make informed decisions about which methods to use in each situation.

B. Structure of this tutorial

This tutorial is organized as follows. We introduce fundamental models of quantum devices in Sec. II, survey common error types in Sec. III, and introduce the most commonly used error metrics in Sec. IV. Once these fundamentals have been introduced, we provide a guide to designing QCVV experiments in Sec. V. A condensed description of preliminary qubit characterization is provided in Sec. VI. Then, we discuss various tomographic QCVV techniques (mostly used to validate and debug small subsystems) in Sec. VII.

We then introduce randomized benchmarks and their theory in Sec. VIII. We conclude this section with a detailed comparison of different benchmarking protocols. In Sec. IX, we discuss "partial tomography" methods that interpolate between randomized benchmarks and full tomographic characterization. In Sec. X, we survey protocols that measure the fidelity of entire quantum circuits. Finally, in Sec. XI, we introduce and survey holistic benchmarks.

II. MODELS OF IMPERFECT QUANTUM COMPUTERS

Real-world quantum computers are complex integrated devices, and the purpose of QCVV experiments is to help understand and predict their behavior. Mathematical models that capture the most salient and important features of a real-world device play an important role in this process. They are essential for *predicting* future behavior (e.g., what will happen when a novel quantum program is run on the device), and highly useful for classifying and understanding the failure modes of quantum computers.

Quantum computers have many subsystems, including control hardware (lasers, arbitrary waveform generators, etc.), environmental management (vacuum chambers, cryostats, shielding, etc.), and more. But at the heart of any gate-based quantum computer is a quantum data register (e.g., an array of qubits) that serves as a physical instantiation of quantum logic and quantum algorithms. Ultimately, the quantum computer's performance can be characterized entirely in terms of this register and the accuracy with which it carries out the quantum logic operations specified by a user. Characterizing and/or benchmarking a quantum computer almost always means probing the behavior of its quantum data register, and other subsystems' behavior is only relevant inasmuch as it impacts the quantum data register.

Thus, the models that underlie characterization and benchmarking protocols must (at a minimum) describe the state of quantum data registers, the actions of quantum logic operations, and the results of measurements. When they function perfectly, relatively simple models suffice. But real quantum registers experience errors that cannot be described by the simplest models. Modeling these errors demands greater accuracy and expressiveness, which requires more complex models. The most commonly used models for qubits and quantum registers fall into three broad categories:

(a) *The closed quantum system model* (Sec. II A). A quantum system that does not interact with its environment evolves reversibly, and is called *closed*. When modeling a closed system, its quantum state is represented by a ray or vector in a *Hilbert space*, a measurement is represented by a *projection-valued measure* (PVM), and an operation on the system (e.g., its dynamical evolution) is represented by a *unitary operator*.

(b) *The Markovian open quantum system model* (Sec. II B). Real-world quantum systems experience irreversible noise when they interact with their environments, and are called *open*. We make the (artificial but useful) assumption that the environment's effects are *Markovian*. In this framework, an open quantum system's state is represented by a *density matrix* on its Hilbert space, a terminating measurement is represented by a *positive operator-valued measure* (POVM), and an operation is represented by a *completely positive trace-preserving* (CPTP) map.

(c) *Non-Markovian open quantum system models*. The category of *non-Markovian* errors includes

an enormous number of effects, including time-correlated noise and coherent coupling to a persistent environment (see Sec. III G for a brief overview). Accurate modeling of quantum systems that experience significant non-Markovian errors typically requires the use of bespoke models that are out of scope for this tutorial.

Because QCVV is primarily concerned with noise and errors, this tutorial uses the Markovian open system model extensively. Representing quantum states as density matrices is straightforward, but the standard models of operations and measurements can get complicated. To lay the necessary groundwork for explaining QCVV metrics and protocols later in this tutorial, we explore three specific topics in detail:

(a) *Representations of quantum operations* (Sec. II C). Quantum operations (e.g., gates) are represented by CPTP linear *superoperators* that map density matrices to density matrices. Several useful and distinct representations of these CPTP maps are used in QCVV.

(b) *Models of quantum measurements* (Sec. II D). Quantum measurements that occur at the end of quantum circuits are called *terminating measurements* and can be modeled by POVMs. To model the internal dynamics of a measurement, or *mid-circuit measurements* that are followed by more gates, more sophisticated models of measurement are needed. Mid-circuit measurements are represented by *quantum instruments*, and if a measurement is weak and perturbs the quantum system minimally, it is possible to continuously track the trajectory of the quantum system in time.

(c) *Gate set models of quantum computers* (Sec. II E). The entire interface of a gate-based quantum computer can be described by a *gate set* that combines models of (i) state preparation, (ii) measurement, and (iii) reversible logic operations. But gate set models are more (or less) than the sum of their parts, because they have *gauge symmetries* that create complications for QCVV.

A. The closed quantum system model

A quantum register is called *closed* if it does not experience noise or interact with any outside systems (i.e., its *environment*). This is an artificial and oversimplified paradigm, but it is simple and elegant, and it is the basis for every introductory quantum mechanics course. Perhaps more importantly, it is the foundation upon which the more complicated and flexible "open quantum system" model is built. We therefore begin by laying out this foundational model, emphasizing the *structure* (mutually consistent mathematical models for quantum states, measurements, and operations) that will be mirrored in the theory of open quantum systems.

1. Quantum state vectors and Hilbert spaces

We can represent the state of a closed quantum register by a *state vector* ψ in the d-dimensional complex vector space $\mathbb{C}^d$, for some integer $d > 0$. This space is denoted $\mathcal{H}$ and called the register's *Hilbert space* [31], and d is the register's *Hilbert-space dimension*.

If N quantum systems with Hilbert-space dimensions $d_1 \dots d_N$ are considered *together* as a single register, the combined system's state is represented by a vector in the tensor product space $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2} \otimes \dots \otimes \mathbb{C}^{d_N}$, and so its Hilbert-space dimension is $\prod_{i=1}^{N} d_i$. Most quantum registers are composed of *n qubits*. A qubit is a quantum system with $d = 2$, so an n-qubit register has $d = 2^n$. In real-world quantum computers, each qubit is *encoded* into a physical system (whose Hilbert-space dimension is $\gg 2$, e.g., an atom) by selecting two quantum states, labeling them "0" and "1," and carefully confining the physical system's quantum state to the two-dimensional subspace that they span. This is often referred to as a *two-level system approximation*. Quantum registers can also be built from qu*dits* with Hilbert space dimension $d > 2$ (e.g., a qu*trit* has $d = 3$, a qu*quart* has $d = 4$, etc.), but this is less common.

Following Dirac's notation, we use *kets* (e.g., $|\psi\rangle$) to denote state vectors. If we specify any orthonormal basis $\{|0\rangle, |1\rangle, \dots, |d-1\rangle\}$ for $\mathbb{C}^d$, then any state $|\psi\rangle$ can be written uniquely as a linear combination of basis vectors, whose coefficients form a column vector:

$$|\psi\rangle = c_0 |0\rangle + c_1 |1\rangle + \dots + c_{d-1} |d-1\rangle \doteq \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{d-1} \end{pmatrix}. \tag{1}$$

We use the $\doteq$ symbol, following Sakurai (Ref. [32], Eq. 1.73), to indicate "is concretely represented by."

In Dirac's notation, the conjugate transpose of $|\psi\rangle$ is a *bra* $\langle\psi|$, whose coefficients form a row vector,

$$\begin{aligned} \langle\psi| &= c_0^* \langle 0| + c_1^* \langle 1| + \dots + c_{d-1}^* \langle d-1|, \\ &\doteq \begin{pmatrix} c_0^* & c_1^* & \cdots & c_{d-1}^* \end{pmatrix}, \end{aligned} \tag{2}$$

and the inner product between two state vectors $|\psi\rangle$ and $|\phi\rangle$ is denoted $\langle\psi|\phi\rangle$. The inner product of $|\psi\rangle$ with itself defines its *norm*, and quantum state vectors are *normalized*:

$$\langle\psi|\psi\rangle = \sum_{i=0}^{d-1} c_i^* c_i = 1. \tag{3}$$

2. Quantum measurements and projection-valued measures

To observe and learn about a quantum system, we perform a *measurement* on it. Many different measurements can be performed on a given system. Measuring yields a particular *outcome*, drawn from a set of d possible outcomes for that measurement. Which outcome occurs is (usually) random, and governed by a probability distribution over the possible outcomes, $\{p(i) : i = 0 \ldots d-1\}$, which is determined by the system's quantum state. The entire purpose of the quantum state is to describe and determine the probabilities of various measurement outcomes, and it is sometimes said that *quantum states are linear functionals on observables*.

Each outcome of a measurement on a closed quantum system is represented by a bra [33] or row vector $\langle\lambda|$. If the measured system is described by state $|\psi\rangle$, then the probability of an outcome labeled "i" represented by $\langle\lambda_i|$ is given by *Born's rule*:

$$p(i|\psi) = |\langle\lambda_i|\psi\rangle|^2 . \tag{4}$$

A measurement is represented by a set of bras that form an orthogonal basis, $\{\langle\lambda_i|\}_{i=0}^{d-1} = \{\langle\lambda_0|, \langle\lambda_1|, \ldots, \langle\lambda_{d-1}|\}$. The corresponding probabilities, $p(i|\psi)$, are all non-negative and add up to 1 because $|\psi\rangle$ is normalized, and thus define a valid probability distribution.

It is clear from Eq. (4) that the state vectors $|\psi\rangle$ and $e^{i\phi}|\psi\rangle$ yield exactly the same probabilities for every measurement. They are, therefore, absolutely indistinguishable by any means, and are considered to define precisely the same state. Here, $e^{i\phi}$ is called a *global phase*, and represents a gauge freedom [34] of this model. However, we can rewrite Born's rule in a way that is useful, suggestive, and eliminates the global phase:

$$p(i|\psi) = \langle\psi|\lambda_i\rangle\langle\lambda_i|\psi\rangle , \tag{5}$$

$$= \mathrm{Tr}[|\lambda_i\rangle\langle\lambda_i|\,|\psi\rangle\langle\psi|] . \tag{6}$$

In this expression, both the state and the measurement outcome are represented as *projectors* (i.e., projection operators) rather than vectors. The global phase freedom vanishes, because $|\psi\rangle\langle\psi|$ is invariant under $|\psi\rangle \to e^{i\phi}|\psi\rangle$, and we have an expression that is *linear* in both $|\psi\rangle\langle\psi|$ and $|\lambda_i\rangle\langle\lambda_i|$. This linearity is extremely useful, and this form of Born's rule motivates the way that both states and measurements are represented for open (noisy) quantum systems.

If we represent the outcomes of a measurement by projectors $|\lambda_i\rangle\langle\lambda_i|$, then the measurement itself is represented by a *set* of mutually orthogonal projectors

$$\{\Pi_0, \Pi_1, \ldots, \Pi_{d-1}\} \equiv \{|\lambda_0\rangle\langle\lambda_0|, |\lambda_1\rangle\langle\lambda_1|, \ldots, |\lambda_{d-1}\rangle\langle\lambda_{d-1}|\} \tag{7}$$

that satisfy *completeness* and *mutual orthogonality* conditions:

$$\sum_i \Pi_i = \mathbb{I} , \tag{8}$$

$$\Pi_i \Pi_j = \delta_{ij} \Pi_i , \tag{9}$$

where δ_{ij} is the Kronecker delta. This set satisfies the mathematical definition of a *measure* (over the set of possible measurement outcomes), and is called a *projection-valued measure (PVM)* [35].

In this model, measurements are assumed to be *repeatable*. Performing a measurement on a system does not destroy it—the system still has a state afterward, and can be measured again—but if the same measurement is performed again, the same outcome will be observed. This requires and implies that if a quantum system is measured and the outcome corresponding to $\langle\lambda_i|$ (or $|\lambda_i\rangle\langle\lambda_i|$) is observed, then its *postmeasurement state* must be $|\lambda_i\rangle$ (or $|\lambda_i\rangle\langle\lambda_i|$):

$$|\psi\rangle \mapsto |\psi'\rangle = |\lambda_i\rangle . \tag{10}$$

Observable properties of a system—e.g., the number of electrons in a quantum dot, or an atom's angular momentum along a particular axis—are represented in this theory by Hermitian operators (acting on the system's Hilbert space) called *observables*. Observables can be measured. Measuring an observable O means performing the PVM whose elements are the projectors onto O's eigenvectors, and indexed by the corresponding eigenvalues of O. So if

$$O = \sum_i o_i |\lambda_i\rangle\langle\lambda_i| , \tag{11}$$

then measuring O on a system in state $|\psi\rangle$ yields value o_i with probability $p(i) = |\langle\psi|\lambda_i\rangle|^2$. The *expectation value* of O is thus

$$\langle O\rangle = \sum_i p(i) o_i = \langle\psi| O |\psi\rangle . \tag{12}$$

3. Qubit state vectors

A qubit is a physical quantum system whose state vector $|\psi\rangle$ is restricted to a two-dimensional subspace $\mathbb{C}^2$ of its Hilbert space. Its state space is the span of two orthogonal *computational basis states*, denoted $|0\rangle$ and $|1\rangle$. They are usually eigenstates of the system's Hamiltonian, and may correspond to an atom's ground and excited states, a photon's horizontal and vertical polarization states, an electron's spin-up and spin-down states, or many other possibilities. Regardless of the physical origin, quantum computation is performed by encoding, manipulating, and measuring these states and/or superpositions of them.

The states $\{|0\rangle, |1\rangle\}$ form a basis for $\mathcal{H}$, so an arbitrary qubit state can be written as a linear combination of them with complex coefficients α and β

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \doteq \begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \tag{13}$$

where the computational basis states themselves are represented by unit column vectors,

$$|0\rangle \doteq \begin{pmatrix}1\\ 0\end{pmatrix}, \quad |1\rangle \doteq \begin{pmatrix}0\\ 1\end{pmatrix}. \tag{14}$$

Because states must be normalized, $\langle\psi|\psi\rangle = |\alpha|^2 + |\beta|^2 = 1$.

Qubit states can be represented as linear combinations of any set of orthonormal basis vectors. Among the most commonly encountered bases are the eigenvectors of the ubiquitous *Pauli operators*. The Pauli operators are represented in the computational basis by 2×2 matrices:

$$\mathbb{I} \doteq \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \tag{15}$$

$$\sigma_x \doteq \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \tag{16}$$

$$\sigma_y \doteq \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}, \tag{17}$$

$$\sigma_z \doteq \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{18}$$

It is common to also use the notation $\{I, X, Y, Z\}$ to denote the Pauli operators. The eigenbasis of the σ_z operator is the computational basis $\{|0\rangle, |1\rangle\}$. The eigenvectors of the σ_x operator, denoted $\{|+\rangle, |-\rangle\}$, can be written as linear combinations of $|0\rangle$ and $|1\rangle$ with real coefficients of equal magnitude $\alpha = \beta = \frac{1}{\sqrt{2}}$:

$$|+\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \quad |-\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}}. \tag{19}$$

Similarly, we can also write $|0\rangle$ and $|1\rangle$ as linear combinations of $|+\rangle$ and $|-\rangle$:

$$|0\rangle = \frac{|+\rangle + |-\rangle}{\sqrt{2}}, \quad |1\rangle = \frac{|+\rangle - |-\rangle}{\sqrt{2}}. \tag{20}$$

Thus, an arbitrary state $|\psi\rangle$ can be expressed in the $\{|+\rangle, |-\rangle\}$ basis as

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle, \tag{21}$$

$$= \alpha\frac{|+\rangle + |-\rangle}{\sqrt{2}} + \beta\frac{|+\rangle - |-\rangle}{\sqrt{2}}, \tag{22}$$

$$= \frac{\alpha+\beta}{\sqrt{2}}\,|+\rangle + \frac{\alpha-\beta}{\sqrt{2}}\,|-\rangle. \tag{23}$$

This is the *same* state as Eq. (13), just written using different basis states.

Since two vectors that differ only by an overall phase describe the same quantum state, we can choose α to be real and, without loss of generality, define $\alpha \equiv \cos(\theta/2)$ and $\beta \equiv e^{i\phi}\sin(\theta/2)$ for some $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. Now, an arbitrary state vector $|\psi\rangle$ is written as

$$|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle. \tag{24}$$

If we compute the expectation values of the three non-identity Pauli operators $\{\sigma_x, \sigma_y, \sigma_z\}$ for this state, we find that

$$\langle\sigma_x\rangle = \sin(\theta)\cos\phi, \tag{25}$$

$$\langle\sigma_y\rangle = \sin(\theta)\sin\phi, \tag{26}$$

$$\langle\sigma_z\rangle = \cos(\theta). \tag{27}$$

This parameterization of the state vector $|\psi\rangle$ associates each qubit state with a unique point on the surface of a unit sphere in $\mathbb{R}^3$, whose coordinates are $\mathbf{r}_\psi = \langle\sigma_x\rangle, \langle\sigma_y\rangle, \langle\sigma_z\rangle$. This is known as the *Bloch sphere*, and the angles θ and ϕ are spherical coordinates for it [see Fig. 1(a)]. The Bloch sphere provides an intuitive visual representation of quantum states and the action of quantum operations on those states, because unitary dynamical evolution (see below) corresponds to rigid rotations of the Bloch sphere. Measuring a qubit whose state is $|\psi\rangle$ in the computational basis will yield a 1-bit result that is "0" with probability $|\alpha|^2 = \cos^2(\theta/2) = (1 + \langle\sigma_z\rangle)/2$ and "1" with probability $|\beta|^2 = \sin^2(\theta/2) = (1 - \langle\sigma_z\rangle)/2$.

4. Dynamical evolution of quantum states

The most important parts of a quantum computation are the dynamical *operations*—e.g., logic gates—performed on the quantum register after it is initialized (in some state) and before it is read out (by measuring it). A quantum operation is a controlled dynamical transformation or evolution of the register's state. Every dynamical evolution of a closed system's state $|\psi\rangle$ is represented by some unitary linear operator U:

$$U : |\psi\rangle \mapsto |\psi'\rangle = U|\psi\rangle. \tag{28}$$

This transformation preserves the state's norm, $\langle\psi|\psi\rangle = \langle\psi'|\psi'\rangle = 1$, and it is reversible because any unitary U has a unitary inverse $U^\dagger = U^{-1}$.

We can choose to model a quantum register like a computer, with a discrete clock cycle. In this paradigm, time takes integer values. In each clock cycle, as time advances from $t-1$ to t, the register's state is transformed by some

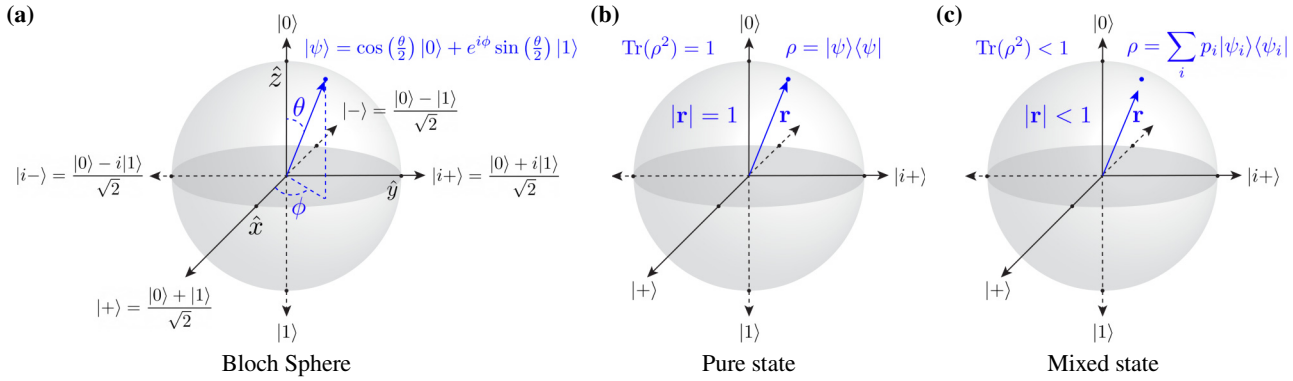


FIG. 1. Bloch ball representation of qubit states. Single-qubit quantum states, usually represented by state vectors $|\psi\rangle$ or density matrices ρ, can also be usefully represented by three-dimensional real vectors called *Bloch vectors* $\mathbf{r}$. Collectively, the Bloch vectors for all qubit states form the *Bloch ball*. Its surface, the *Bloch sphere*, contains the Bloch vectors for all pure qubit states. (a) The computational $|0\rangle$ and $|1\rangle$ states are represented by Bloch vectors at the north and south poles, respectively. Often, these correspond to a system's lowest and second-lowest energy eigenstates. The $|+\rangle$ and $|-\rangle$ states, uniform superpositions of $|0\rangle$ and $|1\rangle$ with real relative phases, are represented by Bloch vectors located along the x axis. The $|i+\rangle$ and $|i-\rangle$ states, uniform superpositions of $|0\rangle$ and $|1\rangle$ with imaginary relative phases, are represented by Bloch vectors located along the y axis. Any pure state can be parameterized as $|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle$ (blue), where θ and ϕ are the polar and azimuthal angles (respectively) of its Bloch vector. (b) Every pure state's Bloch vector has length $|\mathbf{r}| = 1$. (c) The length of the Bloch vector representing a density matrix ρ is $|\mathbf{r}| = \sqrt{2\mathrm{Tr}\rho^2 - 1}$, so Bloch vectors for mixed states have length $|\mathbf{r}| < 1$. In (a)–(c), the blue dot is the projector of the end of the vector onto the surface of the Bloch sphere.

unitary U_t. If we denote the register's state at time t by $|\psi_t\rangle$, then

$$|\psi_1\rangle = U_1 |\psi_0\rangle \ , \tag{29}$$

$$|\psi_2\rangle = U_2 |\psi_1\rangle = U_2 U_1 |\psi_0\rangle \ , \tag{30}$$

and so on. Different gates (or *circuit layers* of parallel gates on distinct parts of the register) will be represented by different unitaries U.

To go deeper and describe a register's detailed dynamics between clock ticks, we can use a continuous-time paradigm in which time t is real-valued. The register's state obeys a differential equation called the *time-dependent Schrödinger equation*,

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = H(t) |\psi(t)\rangle \ , \tag{31}$$

where $H(t)$ is a Hermitian operator called the *Hamiltonian* of the register. It is said to *generate* the register's dynamical evolution in time, which is

$$|\psi(t)\rangle = U(t) |\psi(0)\rangle \tag{32}$$

for some time-dependent $U(t)$ that solves Eq. (31). In the special but useful case where $H(t) = H$ is independent of time, the solution is

$$U(t) = e^{-iHt/\hbar} \, . \tag{33}$$

For arbitrary time-dependent $H(t)$, closed-form solutions to the Schrödinger equation do not generally exist, but many useful approximations and numerical techniques can be used.

B. The Markovian open quantum system model

We assume that the state of a closed quantum system is known as precisely as it can be. This maximal knowledge is represented by a vector $|\psi\rangle$ in Hilbert space. But to model open systems, we need an efficient way to describe states of partial knowledge—e.g., "The system is described by $|\psi_1\rangle$ with probability p_1, and by $|\psi_2\rangle$ with probability p_2." This scenario is not the same as (or consistent with) a superposition state of the form $\alpha|\psi_1\rangle + \beta|\psi_2\rangle$. Instead, it means that in fact the system is either described by $|\psi_1\rangle$ or it is described by $|\psi_2\rangle$, but we are not sure which is true. In such a scenario, we call our description of the system a *mixed state*, to distinguish it from scenarios consistent with a single unique $|\psi\rangle$, which we call a *pure state*.

Mixed states are typically described by an object called a *density matrix*. This new mathematical model for quantum states is richer than the vector state model used to describe pure states of closed systems. It can represent additional forms of uncertainty about the system that state vectors cannot. The theory of *open quantum system* dynamics is built upon density matrices, describing their evolution over time as well as the results of measurements on them.

The density matrix representation of mixed states is nicely motivated by writing the probability of an event i represented by Π_i [Born's rule, Eq. (6)] as

$$p(i|\psi) = \mathrm{Tr}\left[\Pi_i |\psi\rangle\langle\psi|\right] . \tag{34}$$

It follows that if the system is described by $|\psi_j\rangle$ with probability p_j, then the probability of event i is

$$p(i) = \sum_j p_j p\left(i|\psi_j\right), \tag{35}$$

$$= \sum_j p_j \,\mathrm{Tr}[\Pi_i |\psi_j\rangle\langle\psi_j|], \tag{36}$$

$$= \mathrm{Tr}\left[\Pi_i\left(\sum_j p_j |\psi_j\rangle\langle\psi_j|\right)\right]. \tag{37}$$

So, if we represent a pure state by the projector $|\psi\rangle\langle\psi|$, then we can represent the mixed state corresponding to "$|\psi_j\rangle$ with probability p_j" by a probability-weighted linear combination of these projectors,

$$\rho = \sum_j p_j |\psi_j\rangle\langle\psi_j|, \tag{38}$$

so that any measurement probability is given by

$$p(i|\rho) = \mathrm{Tr}[\Pi_i \rho]. \tag{39}$$

ρ is called a *density matrix* (a.k.a. *density operator*). It is a *complete* description of the mixed state! If two different probability distributions over pure states $|\psi_j\rangle$ have identical averages $\rho = \sum_j p_j |\psi_j\rangle\langle\psi_j|$, then those scenarios predict precisely the same probabilities for every possible measurement on the system, and are in fact the same mixed state. Therefore, we always represent mixed states (of open systems) by density matrices. They enable us to model ignorance and uncertainty above and beyond the minimum amount mandated by quantum theory, and to model how noisy operations on a system create or change that "classical" uncertainty.

1. Density matrix formalism

A density matrix ρ is a linear operator that represents the state of a physical quantum system and can be used to compute probabilities for measurement outcomes via Eq. (39):

$$p(i|\rho) = \mathrm{Tr}[\Pi_i \rho].$$

If a system can be described by the pure state $|\psi\rangle$, then its density matrix is the projector

$$\rho = |\psi\rangle\langle\psi|. \tag{40}$$

Density matrices can describe at least two kinds of *uncertain knowledge* that state vectors cannot. The first is represented by a probability distribution over pure states, a.k.a. an *ensemble*, in which the system's state is $|\psi_j\rangle$ with probability p_j [Eq. (38)]:

$$\rho = \sum_j p_j |\psi_j\rangle\langle\psi_j|.$$

The second occurs when a system (S) is *entangled* with a second "reference" system (R) so that they are jointly described by a pure state $|\Psi\rangle_{S,R}$ that is not equal to any tensor product $|\psi\rangle_S \otimes |\phi\rangle_R$. In this scenario, if a measurement is performed on the principal system S, then the probability of an outcome represented by Π_i is

$$p(i|\Psi) = \mathrm{Tr}[(\Pi_i \otimes \mathbb{I}_R)|\Psi\rangle\langle\Psi|]. \tag{41}$$

By writing $\mathbb{I}_R = \sum_j |j\rangle\langle j|_R$, we can show that this probability does not depend on all of $|\Psi\rangle$, but is determined entirely by S's *reduced density matrix*, ρ_S. Specifically,

$$p(i|\Psi) = \mathrm{Tr}[\Pi_i \rho_S], \tag{42}$$

where ρ_S is defined by a *partial trace* over R,

$$\rho_S = \mathrm{Tr}_R[|\Psi\rangle\langle\Psi|] = \sum_j \left(\mathbb{I}_S \otimes \langle j|_R\right)|\Psi\rangle\langle\Psi|\left(\mathbb{I}_S \otimes |j\rangle_R\right). \tag{43}$$

So, a density matrix can predict measurement probabilities (and thus faithfully represent a system's quantum state) both when that system is described by a distribution over pure states, and when it is known to be entangled with another system. These are both mixed states.

Every density matrix must satisfy two key properties:

(1) normalization: $\mathrm{Tr}(\rho) = 1$, and
(2) positive semidefiniteness: $\rho \geq 0$,

which imply three useful facts:

(3) ρ is Hermitian: $\rho = \rho^\dagger$,
(4) $\mathrm{Tr}(\rho^2) \leq 1$, and
(5) $\rho = \rho^2$ if and only if $\mathrm{Tr}(\rho^2) = 1$.

Properties (1)–(2) enforce the basic laws of probability: Property (1) ensures that the outcome probabilities of any measurement sum to 1, and Property (2) ensures that the probability of any measurement outcome is non-negative. Property (3) follows from Property (2), since every positive semidefinite matrix is also Hermitian. Property (4) is a statement about ρ's *purity*,

$$\gamma \equiv \mathrm{Tr}(\rho^2). \tag{44}$$

The maximum possible purity is $\gamma = 1$, achieved uniquely when ρ is a pure state [Eq. (40)]. For any mixed state

[Eq. (38)], $\gamma < 1$, with the minimum possible purity for a d-dimensional state being $\gamma = 1/d$, achieved by the *maximally mixed state* $\rho_{\text{mix}} = \mathbb{I}/d$. Finally, Property (5) follows from Property (4) for pure states; in other words, ρ is *idempotent* if and only if it is pure. Equivalently, if ρ is a projection operator, then it must represent a pure state.

An arbitrary single-qubit pure state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is represented by the density matrix

$$\rho = |\psi\rangle\langle\psi| \doteq \begin{pmatrix} |\alpha|^2 & \alpha\beta^* \\ \alpha^*\beta & |\beta|^2 \end{pmatrix} . \tag{45}$$

If we write $|\psi\rangle$ in spherical coordinates using Eq. (24), we get

$$\rho \doteq \begin{pmatrix} \cos^2\left(\frac{\theta}{2}\right) & e^{-i\phi}\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right) \\ e^{i\phi}\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right) & \sin^2\left(\frac{\theta}{2}\right) \end{pmatrix} , \tag{46}$$

$$= \frac{1}{2}\begin{pmatrix} 1+\cos(\theta) & e^{-i\phi}\sin(\theta) \\ e^{i\phi}\sin(\theta) & 1-\cos(\theta) \end{pmatrix} , \tag{47}$$

$$= \frac{1}{2}\left(\mathbb{I} + \sin(\theta)\cos(\phi)\sigma_x + \sin(\theta)\sin(\phi)\sigma_y + \cos(\theta)\sigma_z\right) . \tag{48}$$

The last expression illustrates a very useful fact: operators form a *vector space*. They can be added, subtracted, and scaled. Any operator can be written as a linear combination of the elements of an *operator basis*. The four Pauli operators form such a basis for operators on qubits, and so we can expand ρ as a linear combination of them. The space of operators on a system's Hilbert space is called its *Hilbert-Schmidt space*, and is used extensively in QCVV. If the system's Hilbert space is denoted $\mathcal{H}$, then its Hilbert-Schmidt space is denoted $\mathcal{B}(\mathcal{H})$ [36]. The inner product between two operators A and B in $\mathcal{B}(\mathcal{H})$ is defined by

$$A \cdot B = \text{Tr}[A^\dagger B] . \tag{49}$$

Operators thus have multiple algebraic roles. They can act as transformations (on the Hilbert space $\mathcal{H}$ of state vectors), or as vectors [in the Hilbert-Schmidt space $\mathcal{B}(\mathcal{H})$ of operators]. When we want to emphasize an operator's vector role, we write it in a double ket or *superket*, e.g., as $|A\rangle\!\rangle$ instead of A. To denote the Hilbert-Schmidt inner product in this notation, we define the *superbra* $\langle\!\langle A| \equiv |A\rangle\!\rangle^\dagger$, so that

$$\langle\!\langle A \,|\, B\rangle\!\rangle = \text{Tr}[A^\dagger B] . \tag{50}$$

The inner product between a density matrix ρ and an observable P is therefore equal to the expectation value of P:

$$\rho \cdot P = \langle\!\langle \rho \,|\, P\rangle\!\rangle = \text{Tr}[\rho P] = \langle P\rangle_\rho , \tag{51}$$

taking advantage of Property (3). We can expand ρ in an orthogonal basis of Hermitian operators $\{P_j\}$ as $\rho = \sum_j c_j P_j$, where each coefficient is given by

$$c_j = \frac{\langle\!\langle \rho \,|\, P_j\rangle\!\rangle}{\langle\!\langle P_j \,|\, P_j\rangle\!\rangle} = \frac{\langle P\rangle_\rho}{\text{Tr}(P_j^2)} . \tag{52}$$

As a result, Eq. (48) (which expands a pure state ρ in the Pauli basis) follows from Eqs. (25)–(27) (the expectation values of the Pauli operators for that state). Therefore, any single-qubit density matrix ρ can be written as

$$\rho = \frac{1}{2}\left(\mathbb{I} + \mathbf{r}\cdot\boldsymbol{\sigma}\right) , \tag{53}$$

where $\mathbf{r} = r_x\hat{\mathbf{x}} + r_y\hat{\mathbf{y}} + r_z\hat{\mathbf{z}}$ and $\boldsymbol{\sigma} = \sigma_x\hat{\mathbf{x}} + \sigma_y\hat{\mathbf{y}} + \sigma_z\hat{\mathbf{z}}$. When $\rho = |\psi\rangle\langle\psi|$ is a pure state, $\mathbf{r}$ is a unit vector. This is the *Bloch sphere* representation mentioned previously, and illustrated in Fig. 1(b). The Bloch-sphere representation of two-level systems is extremely useful for visualizing qubit states prior to measurement and, as we will see in the next chapter, for visualizing the impact of errors on qubits.

A mixed state ρ also defines a vector $\mathbf{r}$, but one with length $|\mathbf{r}|$ less than 1. The length of a mixed state's Bloch vector is determined by its purity:

$$|\mathbf{r}| = \sqrt{2\text{Tr}(\rho^2) - 1} . \tag{54}$$

The $\mathbf{r}$ vectors for mixed states define the *Bloch ball* (the convex closure of the Bloch sphere) as shown in Fig. 1(c), with the maximally mixed state $\rho_{\text{mix}} = \mathbb{I}/2$ at its center ($\mathbf{r} = 0$).

2. Positive operator-valued measures

In closed-system quantum mechanics, a quantum state is represented by a state vector and a measurement is represented by a set of orthogonal projectors (a PVM, as described in Sec. II A 2). In open quantum systems, we need to model additional uncertainty. This requires richer representations not just of states (as density matrices), but of measurements as well. If S is an open quantum system, then it is possible to perform *indirect* measurements on S by (1) coupling S to another system R [Fig. 2(a)], and then (2) performing a PVM on S and R jointly [Fig. 2(b)]. This enables and allows a substantially richer class of measurements called a *positive operator-valued measure (POVM)* [37].

A POVM is a set of positive semi-definite operators $\{E_i\}$ that satisfies the completeness relation:

$$\sum_i E_i = \mathbb{I} . \tag{55}$$

Each E_i is called an *effect*, and represents one possible outcome "i" of the measurement. When the POVM $\{E_i\}$

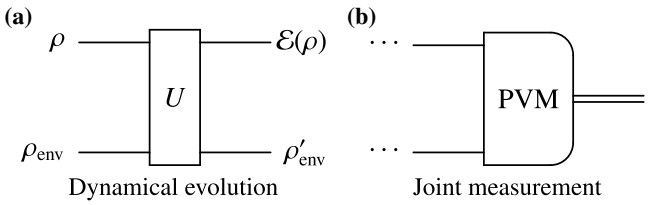


FIG. 2. System-environment representation of quantum operations. (a) An open quantum system comprises a principal system of interest whose state is ρ, and a surrounding environment whose state is ρ_{env}. Any dynamical evolution of the principal system can be described by some joint evolution of the system and its environment, described by a unitary transformation U that may produce correlation or entanglement between system and environment. After this evolution, the state of the principal system can be described by a reduced density matrix, defined as a partial trace over the environment. This density matrix is a linear function $\mathcal{E}(\rho)$ of the initial state ρ. (b) Terminating measurements of the principal system $\mathcal{E}(\rho)$ can be modeled by performing a PVM on the joint system $[\mathcal{E}(\rho) \otimes \rho'_{\text{env}}]$. The results of this joint measurement are described by POVMs, whose outcome is a distribution of classical bits (denoted by the double line).

is performed on a state ρ, the probability of observing outcome i is given by

$$p(i|\rho) = \text{Tr}[E_i \rho], \tag{56}$$

which is the open-system version of Born's rule.

Any projective measurement (PVM) is also a POVM. But POVMs are quite a bit more general. The effects in a POVM do not need to be orthogonal, nor rank-1, nor projectors. *Any* set satisfying the conditions above is a valid, feasible POVM. Importantly, POVMs describe *destructive* measurements, and do not specify or define what a system's state will be after measurement.

3. Dynamical evolution of density matrices

An open quantum system can interact with its environment. This possibility allows new kinds of dynamical evolution that can create or decrease uncertainty, causing the open system's state to become more (or less) mixed. In contrast, closed-system dynamics are always unitary and never change the purity of ρ. An open-system dynamical evolution $\rho \mapsto \mathcal{E}(\rho)$ can be described by a three-step process, illustrated in Fig. 2(a):

(1) The principal system of interest is described (initially) by a state ρ. We introduce a second system, the environment, that is assumed to be initially uncorrelated with the principal system [38] and described by its own state $\rho_{\text{env}} = |e_0\rangle\langle e_0|$.
(2) The system and environment evolve jointly, according to familiar closed-system theory, by some unitary U that may induce correlations or entanglement between them.
(3) We focus on the principal system only, neglecting or "throwing away" the environment by performing a partial trace over its Hilbert space, to obtain a reduced density matrix for the principal system only:

$$\mathcal{E}(\rho) = \text{Tr}_{\text{env}}[U(\rho \otimes \rho_{\text{env}})U^\dagger]. \tag{57}$$

The dynamical map $\rho \mapsto \mathcal{E}(\rho)$ is called a *quantum operation* (a.k.a. a *quantum channel*), and Eq. (57) is known as the *system-environment*, or *Stinespring*, representation of quantum operations; it is depicted graphically in Table I. It stems from Stinespring's *dilation theorem* [39], which states that every physically allowed dynamical evolution of the principal system arises from unitary evolution on a larger system, and can thus be described by Eq. (57). In terms of the nonsquare *Stinespring operator* defined as $A \equiv U(\mathbb{I} \otimes |e_0\rangle)$, Eq. (57) is simply

$$\mathcal{E}(\rho) = \text{Tr}_{\text{env}}[A\rho A^\dagger]. \tag{58}$$

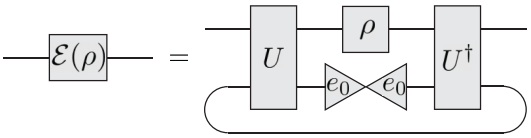
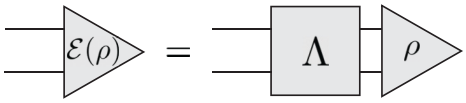
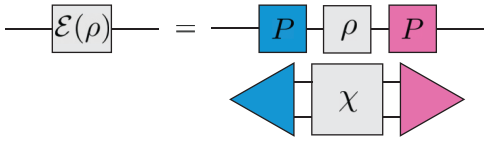
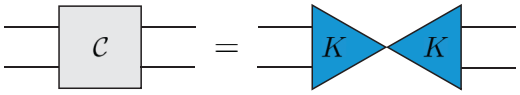
Equation (58) shows clearly that the dynamical map $\mathcal{E}$ acts *linearly* on density matrices—i.e., if $\rho = a\rho_1 + b\rho_2$, then $\mathcal{E}(\rho) = a\mathcal{E}(\rho_1) + b\mathcal{E}(\rho_2)$.

Such a linear map can represent a real, physically realizable quantum operation if—and only if—it satisfies two conditions:

(1) *Complete positivity (CP)*: Given any positive semidefinite density matrix $\rho \geq 0$, applying $\mathcal{E}$ to ρ must yield a matrix that is also positive semidefinite, even if $\mathcal{E}$ only acts on a part (subsystem) of ρ. So $(\mathcal{E} \otimes \mathbb{I})[\rho] \geq 0$ for every $\rho \geq 0$. This is a stricter requirement than simple *positivity*—$\mathcal{E}[\rho] \geq 0$ for every $\rho \geq 0$—because a linear map can be positive yet not completely positive. A canonical example is the transpose map, $\mathcal{E}(\rho) = \rho^{\text{T}}$. If such a map could be experimentally applied to arbitrary states, then by applying it to a system properly entangled with another system, a negative probability (for some measurement outcome) could be produced.
(2) *Trace preservation (TP)*: $\text{Tr}[\mathcal{E}(\rho)] = \text{Tr}[\rho]$ for all ρ.

Just like Properties (1)–(2) of density matrices, these conditions guarantee the two essential properties of probability distributions. CP ensures that no event can have negative probability, while TP ensures that the outcome probabilities of every measurement add up to 1. Linear maps satisfying both conditions are called *completely positive trace-preserving (CPTP)* maps, and every CPTP map represents a physically realizable quantum operation.

TABLE I. Graphical representations of operations. Reference [40] introduces a useful and intuitive "graphical calculus," closely related to the visual language of quantum circuits. These graphical representations can be a useful aid to intuition for students learning the mathematics of open quantum systems. This table shows the graphical representations of the Stinespring, Kraus, transfer matrix, Chi (process) matrix, and Choi matrix representations of a quantum operation. Although we do not explain here how these diagrams are interpreted and manipulated, the interested reader is referred to Ref. [40] for a comprehensive discussion of graphical calculus and its use in quantum computing.

Mathematical	Graphical	Representation
$\mathcal{E}(\rho) = \mathrm{Tr}_{\mathrm{env}}[U(\rho \otimes \rho_{\mathrm{env}})\, U^\dagger]$		Stinespring
$\mathcal{E}(\rho) = \sum_i^N K_i \rho K_i^\dagger$		Kraus
$\|\mathcal{E}(\rho)\rangle\rangle = \Lambda \|\rho\rangle\rangle$		Transfer matrix
$\mathcal{E}(\rho) = \sum_{jk} \chi_{jk} P_j \rho P_k$		Chi (process) matrix
$\mathcal{C} = \sum_k \mathbf{vec}(K_k)\mathbf{vec}(K_k)^\dagger$		Choi matrix

The Stinespring representation of an operation $\mathcal{E}$ is not unique, because many different physical scenarios (U and ρ_{env}) can produce identical reduced dynamics for the principal system. As a result, the Stinespring representation is rarely used in practical calculations, because more convenient representations exist. As we will see in this tutorial, quantum operations are very important in QCVV, and the QCVV literature uses several distinct representations of them for specific purposes. We examine these representations in detail in the next subsection.

C. Representations of quantum operations

We have already seen one way to represent a quantum operation $\mathcal{E}$, as a unitary transformation on a larger Hilbert space [Eq. (57)]. In this section, we will construct several more, each with their own unique properties and advantages. These mathematical representations can also be represented *graphically*, which the interested reader might find helpful for understanding the math of this section. We provide several examples of these graphical representations in Table I, but a formal introduction to this material is beyond the scope of this tutorial. For a complete introduction to the graphical calculus of open quantum systems, the reader is referred to Ref. [40].

1. Kraus (operator-sum) representation

We can construct a second representation by rewriting Eq. (57) in terms of a set of orthonormal basis states $\{|e_i\rangle\}$ for the environment's Hilbert space, as

$$\mathcal{E}(\rho) = \sum_i \langle e_i|\, U(\rho \otimes |e_0\rangle\langle e_0|)\, U^\dagger\, |e_i\rangle\ . \tag{59}$$

If we define $K_i \equiv \langle e_i|\, U\, |e_0\rangle$, then Eq. (59) becomes

$$\mathcal{E}(\rho) = \sum_i^N K_i \rho K_i^\dagger\ . \tag{60}$$

The operators $\{K_i\}$ are known as *Kraus operators*, and Eq. (60) is known as the *Kraus* or *operator-sum* representation of a quantum operation $\mathcal{E}$; it is depicted graphically in Table I. The Kraus representation is also not unique—distinct sets $\{K_i\}$ and $\{K_i'\}$ can produce identical operations $\mathcal{E}$. However, it is always possible to construct a Kraus representation with $N \leq d^2$ Kraus operators K_i (where $d = 2^n$ for n qubits) that are mutually orthogonal—i.e., $\mathrm{Tr}(K_i^\dagger K_j) = 0$ for $i \neq j$. This representation is almost always unique (see Sec. 8.2.4 of Ref. [41] for an in-depth analysis).

The Kraus representation of $\mathcal{E}$ avoids any explicit reference to the environment's state or dynamics, describing

the evolution of ρ using only operators acting on the principal system. The Kraus operators do not need to be derived (as we did above) from properties of the environment. *Any* Kraus representation automatically satisfies the CP condition, and a set of Kraus operators $\{K_i\}$ satisfies the TP condition (and thus describes a physically allowed quantum operation) if and only if it satisfies a *completeness relation*:

$$\sum_i^N K_i^\dagger K_i = \mathbb{I}\,. \tag{61}$$

If the Kraus operators of a single-qubit operation $\mathcal{E}$ are proportional to Pauli operators, then we call $\mathcal{E}$ a *Pauli channel*. This concept can be extended to quantum operations acting on $n > 1$ qubits using the n-qubit Pauli operators, which comprise all 4^n tensor products of one-qubit Pauli operators. If we define $\mathbb{P} = \{I, X, Y, Z\}$, then the n-qubit Pauli group $\mathbb{P}_n$ is given by

$$\mathbb{P}_n \equiv \mathbb{P}^{\otimes n} = \{I, X, Y, Z\}^{\otimes n}\,. \tag{62}$$

An n-qubit operation $\mathcal{E}$ is a Pauli channel if

$$\mathcal{E}(\rho) = \sum_{P\in\mathbb{P}_n} p_P P\rho P^\dagger \tag{63}$$

for some probability distribution $\{p_P\}$. Pauli channels are useful and intuitive because they describe probabilistic (a.k.a., stochastic or random) processes. Each Pauli operator P is a unitary operation that could "happen" to ρ, and if ρ evolves according to a Pauli channel, then P occurs with probability p_P.

2. Transfer matrix representation

The third common representation of a quantum operation $\mathcal{E}$ is as a *linear superoperator* or *transfer matrix*. We saw above that a quantum operation's action on a density matrix must be described by a *linear* map $\rho \mapsto \mathcal{E}(\rho)$. And, as observed previously, density matrices describing a system's state can be thought of as vectors in the Hilbert-Schmidt space of $d \times d$ matrices. So, just as $|\psi\rangle$ is a vector in Hilbert space that is transformed by unitary operators U, ρ can be viewed as a vector $|\rho\rangle\!\rangle$ in Hilbert-Schmidt space that is transformed by linear maps $\mathcal{E}$.

We can make this action explicit by taking a $d \times d$ density matrix ρ whose elements in a standard basis are

$$\rho \doteq \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1d} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d1} & \rho_{d2} & \cdots & \rho_{dd} \end{pmatrix} \tag{64}$$

and *vectorizing* [42] it into a $d^2 \times 1$ column vector [43]

$$|\rho\rangle\!\rangle \doteq (\rho_{11}, \rho_{21}, \ldots, \rho_{d1}, \rho_{12}, \ldots, \rho_{dd})^{\mathrm{T}}\,. \tag{65}$$

This particular representation, sometimes called "column stacking" because the columns of ρ are stacked atop one another, is one specific way to represent $|\rho\rangle\!\rangle$ concretely in a particular operator basis. To get Eq. (65), we use the basis of *matrix units* defined (in terms of the Hilbert space basis used to represent ρ) by $\{|i\rangle\langle j|\,,\ i,j = 1\ldots d\}$, because

$$\rho_{ij} = \mathrm{Tr}[|j\rangle\langle i|\,\rho] = \langle\!\langle |i\rangle\langle j|\,||\rho\rangle\!\rangle\,. \tag{66}$$

But an operator A can be "vectorized" in any operator basis. One particularly useful choice is the n-qubit Pauli basis. However, whereas the matrix units are orthonormal because

$$\langle\!\langle |i\rangle\langle j|\,||\,|k\rangle\langle l|\rangle\!\rangle = \mathrm{Tr}[|j\rangle\langle i||k\rangle\langle l|] = \delta_{ik}\delta_{jl}\,, \tag{67}$$

the Paulis are mutually orthogonal but not normalized, because

$$\langle\!\langle P_i\,|\,P_j\rangle\!\rangle = \delta_{ij}\,\mathrm{Tr}[\mathbb{I}] = d\delta_{ij}\,. \tag{68}$$

This can be dealt with either by using *normalized* Pauli operators $\{P_i/\sqrt{d}\}$, or by computing the coefficient of each Pauli basis operator as $\langle\!\langle P_i\,|\,A\rangle\!\rangle/d$ instead of $\langle\!\langle P_i\,|\,A\rangle\!\rangle$.

Using this framework, a quantum operation $\mathcal{E}$ is just a linear transformation that maps any density matrix $\rho \in \mathcal{B}(\mathcal{H})$ to a new density matrix $\rho' \in \mathcal{B}(\mathcal{H})$ in the same Hilbert-Schmidt space [44]. Therefore, $\mathcal{E}$ can be concretely represented by a $d^2 \times d^2$ matrix Λ that acts on vectorized states $|\rho\rangle\!\rangle$ by matrix multiplication:

$$|\rho\rangle\!\rangle \mapsto |\mathcal{E}(\rho)\rangle\!\rangle = \Lambda|\rho\rangle\!\rangle\,. \tag{69}$$

A linear map on density matrices, or a matrix that acts on vectorized density matrices, is called a *superoperator*. Every quantum operation can be described by a superoperator Λ. Superoperators representing quantum operations are often called *transfer matrices* because Λ's action on vectorized density matrices resembles the action of transfer operators in dynamical systems or statistical mechanics. This representation is also sometimes called the *Liouville* or *associative* representation, and is depicted graphically in Table I.

An operation $\mathcal{E}$'s transfer matrix Λ can be concretely constructed by choosing an orthonormal basis $\{B_i\}$ for the vector space of $d \times d$ matrices, then defining the elements of Λ using the Hilbert-Schmidt inner product:

$$\Lambda_{ij} \equiv \mathrm{Tr}\left[B_i^\dagger \mathcal{E}(B_j)\right]\,. \tag{70}$$

It is common to construct Λ in the basis of matrix units, or the Pauli basis (see Sec. II C 3). We usually work in

the Pauli basis; transfer matrices and other objects that are explicitly represented in the basis of matrix units are identified with the subscript c. If $\mathcal{E}$ has Kraus operators $\{K_i\}$, then its transfer matrix in the basis of matrix units is

$$\Lambda_c = \sum_i K_i^* \otimes K_i \,. \tag{71}$$

In the transfer matrix representation, the composition of quantum operations is associative. In other words, if two operations with transfer matrices Λ_1 and Λ_2 are applied in succession, then the net effect is to apply the product of the two matrices, i.e., $\Lambda_2\Lambda_1$:

$$|\rho'\rangle\rangle = |\mathcal{E}_2(\mathcal{E}_1(\rho))\rangle\rangle = \Lambda_2\,|\mathcal{E}_1(\rho)\rangle\rangle = \Lambda_2\Lambda_1\,|\rho\rangle\rangle\,. \tag{72}$$

Because of this useful property, the transfer matrix representation is widely used to model errors in quantum logic operations (i.e., gates). When those errors are *small*, a variation called the *error generator* formalism [45]—that represents transfer matrices by their logarithms—is useful for distinguishing and classifying small errors (see Appendix A for a brief summary). Error generators can be used to classify and quantify the rates at which different types of errors occur in quantum gates [46].

3. Pauli transfer matrix representation

When a transfer matrix is constructed in the Pauli basis, we call it a *Pauli transfer matrix (PTM)*. This is also sometimes known as the *Pauli-Liouville* representation of quantum operations. We use the symbol Λ for all transfer matrix representations, regardless of basis, but in this tutorial Λ will always indicate a PTM unless another basis is specified.

The PTM representation of a quantum operation $\mathcal{E}$ is a $4^n \times 4^n$ superoperator Λ with entries

$$\Lambda_{ij} = \langle\langle P_i\,|\,\mathcal{E}(P_j)\rangle\rangle = \frac{1}{d}\mathrm{Tr}\left[P_i\mathcal{E}(P_j)\right], \tag{73}$$

where P_i and P_j are elements of the n-qubit Pauli group $\mathbb{P}_n$. PTMs act on density matrices, which have also been expanded in the Pauli basis,

$$\rho = \sum_{P\in\mathbb{P}_n} \rho_P P\,, \tag{74}$$

where $\rho_P = \langle\langle P\,|\,\rho\rangle\rangle/d$ are the expansion coefficients. By vectorizing the expansion coefficients into a single column vector,

$$|\rho\rangle\rangle \doteq \begin{pmatrix}\rho_{I^{\otimes n}} & \cdots & \rho_{Z^{\otimes n}}\end{pmatrix}^{\mathrm{T}}, \tag{75}$$

the quantum map $\rho' = \mathcal{E}(\rho)$ can be expressed in vector form, where the PTM Λ acts on $|\rho\rangle\rangle$ by direct matrix multiplication: $|\rho'\rangle\rangle = \Lambda|\rho\rangle\rangle$. For example, for a single qubit,

$$|\rho\rangle\rangle \doteq \begin{pmatrix}\rho_I\\ \rho_X\\ \rho_Y\\ \rho_Z\end{pmatrix}, \tag{76}$$

and the map $\rho' = \mathcal{E}(\rho)$ is given by

$$\begin{pmatrix}\rho'_I\\ \rho'_X\\ \rho'_Y\\ \rho_Z\end{pmatrix} = \begin{pmatrix}\Lambda_{II} & \Lambda_{IX} & \Lambda_{IY} & \Lambda_{IZ}\\ \Lambda_{XI} & \Lambda_{XX} & \Lambda_{XY} & \Lambda_{XZ}\\ \Lambda_{YI} & \Lambda_{YX} & \Lambda_{YY} & \Lambda_{YZ}\\ \Lambda_{ZI} & \Lambda_{ZX} & \Lambda_{ZY} & \Lambda_{ZZ}\end{pmatrix}\begin{pmatrix}\rho_I\\ \rho_X\\ \rho_Y\\ \rho_Z\end{pmatrix}. \tag{77}$$

The entries of a PTM are all real numbers bounded by $\Lambda_{ij} \in [-1, 1]$. Some important properties of an operation can be extracted directly from its PTM. We can isolate four (slightly overlapping) useful blocks within a PTM, as shown in Fig. 3.

(a) Λ's top row reveals whether it is trace preserving. The operation is TP if and only if $\Lambda_{0j} = \delta_{0j}$ (i.e., if the first row of the PTM is $[1, 0, \ldots, 0]$). Every deterministic process must be TP, but postselected operations provide an example of non-TP processes [47].

(b) The bottom right $(d^2-1)\times(d^2-1)$ block of the PTM is called the *unital* block. An operation is unital if it preserves the identity [i.e., $\mathcal{E}(\mathbb{I}) = \mathbb{I}$], and the PTM for a unital operation is restricted to this block and the 1×1 block defined by Λ_{00}. Unital processes cannot increase purity (or decrease entropy). Unitary dynamics (Sec. III A) and stochastic Pauli errors (Sec. III E) are unital.

(c) The leftmost column of Λ is called the *nonunital* block. For any unital operation, $\Lambda_{i0} = \delta_{i0}$ (i.e., the first column of the PTM is $[1, 0, \ldots, 0]^{\mathrm{T}}$). If it does not take this form, its elements indicate entropy-decreasing processes like cooling, energy relaxation, or spontaneous emission (e.g., T_1 decay; see Sec. III C).

(d) The diagonal elements of Λ quantify how well *polarization* is preserved along each Pauli axis (see Sec. IV C 4), with $\Lambda_{PP} = 1$ if the operation preserves the component of Pauli operator P in ρ. $\Lambda_{PP} < 1$ indicates loss of polarization or coherence along a Pauli axis. A PTM's diagonal elements Λ_{PP} are sometimes called the *survival probabilities*, *Pauli fidelities*, or (when Λ is diagonal) *Pauli eigenvalues*. An operation's PTM is diagonal if and only if it is a Pauli channel.

The TP constraint is obvious and easy to enforce in the PTM representation, by requiring that $\Lambda_{0j} = \delta_{0j}$. In contrast, the CP constraint is hard to express or evaluate in

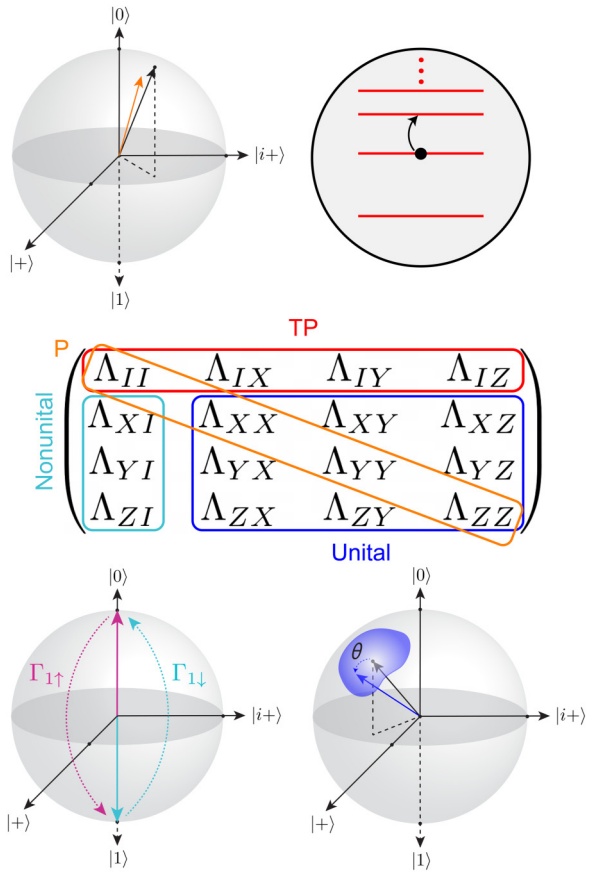


FIG. 3. Pauli transfer matrix. We can identify four useful blocks within a PTM. The top row is typically fixed as $[1, 0, 0, 0]$ by trace preservation (TP, red), although postselected operations can be non-TP. The lower right-hand block (blue) captures unital processes, such as unitary errors. The column to the left of the unital block (cyan) indicates nonunital processes, such as T_1 decay, resulting in $\Lambda_{PI} \neq 0$ for $P \in \{X, Y, Z\}$. The diagonal elements indicate how well polarization (P, orange) along the various Pauli axes is preserved, and are directly impacted by stochastic Pauli noise. The spheres at each corner depict example errors captured by each block (see Fig. 4).

the PTM representation. The easiest way to test whether a PTM Λ describes a CP map is to construct its Choi matrix representation (see Sec. II C 5).

An operation's PTM can be computed directly from a transfer matrix represented in a different basis [Eq. (70)] by applying a unitary change of basis. Suppose, for example, Λ_c is a transfer matrix written in the basis of matrix units [Eq. (71)]. We can construct the equivalent PTM Λ as

$$\Lambda = U_{c\to P}\Lambda_c U_{c\to P}^\dagger\,, \tag{78}$$

where

$$U_{c\to P} \equiv \frac{1}{\sqrt{d}}\sum_{i=0}^{d^2-1} |c_i\rangle\!\rangle\langle P_i|\,, \tag{79}$$

where $\{|c_i\rangle\!\rangle\}_{i=0}^{d^2-1}$ is the basis of matrix units. The factor of $1/\sqrt{d}$ normalizes the Pauli basis elements to 1 and makes $U_{c\to P}$ unitary. The inverse transformation is also possible with

$$\Lambda_c = U_{P\to c}\Lambda U_{P\to c}^\dagger\,, \tag{80}$$

where $U_{P\to c} = U_{c\to P}^\dagger$.

4. Chi (process) matrix representation

We can construct a fourth representation of $\mathcal{E}$ by expanding $\mathcal{E}$'s Kraus operators $\{K_i\}$ [Eq. (60)] in a fixed operator basis $\{P_j\}$, such as the Pauli basis:

$$K_i = \sum_j c_{ij} P_j\,, \tag{81}$$

where c_{ij} are the expansion coefficients. Plugging this expansion into Eq. (60) yields

$$\mathcal{E}(\rho) = \sum_{jk} \chi_{jk} P_j \rho P_k\,, \tag{82}$$

where $\chi_{jk} = \sum_i c_{ij} c_{ik}^*$. The $d^2 \times d^2$ matrix of coefficients χ_{jk} is called a χ *matrix*, and this is known as the χ *matrix representation* of an operation $\mathcal{E}$; it is depicted graphically in Table I. Historically, an operation's χ matrix was called its *process matrix*. More recently, the term "process matrix" has been used more broadly to describe other matrix representations of $\mathcal{E}$ (e.g., the Pauli transfer matrix). In this tutorial, *process matrix* will always refer to the χ matrix, and χ matrices will be constructed in the Pauli basis unless otherwise specified.

The χ matrix representation is closely related to the Kraus representation, but has certain advantages. It is easy to construct mechanically, and it is unique once an operator basis $\{P_j\}$ is chosen. A map $\mathcal{E}$ is CP if and only if its χ matrix is positive semidefinite. Constructing $\mathcal{E}$'s χ matrix and checking whether it is positive semidefinite is the easiest way to check complete positivity. Trace preservation (TP) is also easy to check in this representation; $\mathcal{E}$ is TP if and only if $\sum_{j,k} \chi_{jk} P_j^\dagger P_k = \mathbb{I}$. This condition is equivalent to Eq. (61). Moreover, it constrains d^2 of the d^4 parameters in χ, and so the χ matrix for a CPTP map has $d^2(d^2-1)$ free parameters.

Equation (82) can be seen as an expansion of $\mathcal{E}$ as a linear combination of non-CPTP linear maps known as *Choi units*, denoted X:

$$\mathcal{E} = \sum_{jk} \chi_{jk} X_{jk}\,, \tag{83}$$

$$X_{jk}(\rho) \equiv P_j \rho P_k\,. \tag{84}$$

Each Choi unit is a superoperator acting on operators. Choi units are Hermitian ($X_{jk}^\dagger = X_{jk}$), and they form an

orthogonal basis,

$$\mathrm{Tr}[X_{ij}^\dagger X_{kl}] = d^2\delta_{ik}\delta_{jl}\,. \tag{85}$$

This makes it easy to construct the χ matrix for any operation $\mathcal{E}$, since

$$\chi_{jk} = \frac{1}{d^2}\mathrm{Tr}[X_{jk}^\dagger \mathcal{E}] = \frac{1}{d^3}\sum_{P_m\in\mathbb{P}_n}\mathrm{Tr}[P_m P_j \mathcal{E}(P_m)P_k]\,. \tag{86}$$

For example, the χ matrix coefficients for the identity operation $\mathcal{E}(\rho)=\rho$ are

$$\chi_{jk} = \frac{1}{d^3}\sum_{P_m\in\mathbb{P}_n}\mathrm{Tr}[P_m P_j P_m P_k]\,, \tag{87}$$

$$= \frac{1}{d^2}\mathrm{Tr}[P_j]\mathrm{Tr}[P_k]\,, \tag{88}$$

$$= \delta_{j0}\delta_{k0}\,, \tag{89}$$

using the identity $\sum_{P\in\mathbb{P}_n} PMP = d\mathrm{Tr}(M)\mathbb{I}$.

Any χ matrix for a CP map can be diagonalized by a unitary change of basis. This diagonal form,

$$\mathcal{E}(\rho) = \sum_j \lambda_j Q_j \rho Q_j^\dagger\,, \tag{90}$$

gives the orthogonal Kraus representation of $\mathcal{E}$, with $K_i = \sqrt{\lambda_i}Q_i$. It is unique up to degeneracies ($\lambda_j = \lambda_{j'}$ for some $j\neq j'$). Constructing the χ matrix in some basis and diagonalizing it is the easiest way to find the orthogonal Kraus form of a generic operation.

Both the PTM and χ matrix representations of a map $\mathcal{E}$ are unique. So, it is possible to compute the PTM from the χ matrix as

$$\Lambda_{ij} = \frac{1}{d}\sum_{kl}\chi_{kl}\mathrm{Tr}\left[P_i P_k P_j P_l\right]\,, \tag{91}$$

and the χ matrix from the PTM as

$$\chi_{ij} = \frac{1}{d^3}\sum_{kl}\Lambda_{kl}\mathrm{Tr}\left[P_l P_i P_k P_j\right]\,. \tag{92}$$

5. Choi matrix representation

The final representation of a quantum operation $\mathcal{E}$ that we consider represents $\mathcal{E}$ as an unnormalized quantum state of a larger (two-copy) system. We begin with a maximally entangled state of two systems, which can be defined in terms of any orthonormal basis $\{|i\rangle\}_{i=0}^{d-1}$ as

$$|\Phi_0\rangle = \sum_i |i\rangle\otimes|i\rangle\,, \tag{93}$$

keeping in mind that this "state" has norm d. Its density matrix is

$$|\Phi_0\rangle\langle\Phi_0| = \sum_{i,j}|i\rangle\langle j|\otimes|i\rangle\langle j|\,, \tag{94}$$

which for an n-qubit system can also be written using Pauli operators as

$$|\Phi_0\rangle\langle\Phi_0| = \frac{1}{d}\sum_{P\in\mathbb{P}_n} P\otimes P^{\mathrm{T}}\,. \tag{95}$$

We now apply $\mathbb{I}\otimes\mathcal{E}$ to this state to get

$$\mathcal{C} = (\mathbb{I}\otimes\mathcal{E})\left[|\Phi_0\rangle\langle\Phi_0|\right], \tag{96}$$

$$= \sum_{i,j=0}^{d-1}|i\rangle\langle j|\otimes\mathcal{E}(|i\rangle\langle j|)\,, \tag{97}$$

$$= \frac{1}{d}\sum_{P\in\mathbb{P}_n} P\otimes\mathcal{E}(P)^{\mathrm{T}}\,. \tag{98}$$

$\mathcal{C}$ is called the *Choi matrix* [48], or sometimes the *dynamical matrix* [49], of $\mathcal{E}$. The action of an operation $\mathcal{E}$ can be written in terms of its Choi matrix $\mathcal{C}$ as

$$\mathcal{E}(\rho) = \mathrm{Tr}_1\left[\mathcal{C}(\rho^{\mathrm{T}}\otimes\mathbb{I})\right], \tag{99}$$

where Tr_1 denotes the partial trace over the first subsystem. By substituting this into Eq. (97), it is straightforward to verify that $\mathcal{C}$ faithfully and uniquely represents $\mathcal{E}$.

This one-to-one correspondence between completely positive *operations* $\mathcal{E}$ and positive semidefinite (bipartite) *states* $\mathcal{C}$ is known as the *Choi-Jamiołkowski isomorphism* (or *channel-state duality*) [50,51]. It implies several useful properties:

(1) $\mathcal{E}$ is CP $\iff$ $\mathcal{C}\geq 0$,
(2) $\mathcal{E}$ is TP $\iff$ $\mathrm{Tr}_1[\mathcal{C}]=\mathbb{I}$,
(3) $\mathcal{E}$ is Hermitian-preserving (HP) $\iff$ $\mathcal{C}=\mathcal{C}^\dagger$,

where $\iff$ denotes *if and only if*.

An even more fundamental (and perhaps surprising) statement is that the Choi matrix is proportional to the transpose of the χ matrix in suitably chosen bases. In other words, given any basis $\{|\phi_i\rangle\}$ for the bipartite Hilbert space $\mathcal{H}\otimes\mathcal{H}$, there is a corresponding operator basis $\{|P_i\rangle\rangle\}$

for $\mathcal{B}(\mathcal{H})$ so that the χ and Choi representations of any operation $\mathcal{E}$ obey

$$\chi_{ji} = \frac{1}{d}\langle\phi_i|\mathcal{C}|\phi_j\rangle\,. \tag{100}$$

We can demonstrate this using the Pauli operator basis, and the basis of *normalized* maximally entangled states (for $\mathcal{H}\otimes\mathcal{H}$) given by $\{|\phi_j\rangle\} = \left\{(P_j\otimes\mathbb{I})\,|\Phi_0\rangle\,/\sqrt{d}\right\}_{P_j\in\mathbb{P}_n}$. Using Eqs. (97) and (95), we can write

$$\frac{1}{d}\langle\phi_i|\mathcal{C}|\phi_j\rangle = \frac{1}{d}\mathrm{Tr}[\mathcal{C}(P_j\otimes\mathbb{I})\,|\phi_0\rangle\langle\phi_0|\,(P_i\otimes\mathbb{I})]\,, \tag{101}$$

$$= \frac{1}{d^4}\sum_{k,l}\mathrm{Tr}[(P_k\otimes\mathcal{E}(P_k)^{\mathrm{T}})(P_j\otimes\mathbb{I})(P_l\otimes P_l^{\mathrm{T}}) \times (P_i\otimes\mathbb{I})]\,, \tag{102}$$

$$= \frac{1}{d^4}\sum_{k,l}\mathrm{Tr}[P_kP_jP_lP_i]\mathrm{Tr}[P_l\mathcal{E}(P_k)]\,, \tag{103}$$

$$= \frac{1}{d^3}\sum_{k,l}\Lambda_{lk}\mathrm{Tr}[P_lP_iP_kP_j]\,, \tag{104}$$

$$= \chi_{ji}\,. \tag{105}$$

where in the last line we have used the cyclic property of the trace and Eq. (92). This equivalence is very powerful, since it implies that $\mathcal{C}$ and χ are essentially identical [up to a transpose, a factor of d, and appropriate choice of bases for $\mathcal{H}\otimes\mathcal{H}$ and $\mathcal{B}(\mathcal{H})$].

If an operation $\mathcal{E}$ has the Kraus representation $\mathcal{E}(\rho) = \sum_k K_k\rho K_k^\dagger$, then we can construct its Choi representation (depicted graphically in Table I) beginning with Eq. (97) as

$$\mathcal{C} = \sum_{i,j=0}^{d-1}|i\rangle\langle j|\otimes\mathcal{E}(|i\rangle\langle j|)\,, \tag{106}$$

$$= \sum_{i,j,k}|i\rangle\langle j|\otimes K_k\,|i\rangle\langle j|\,K_k^\dagger\,, \tag{107}$$

$$= \sum_{i,j,k}\left(|i\rangle\otimes K_k\,|i\rangle\right)\left(\langle j|\otimes\langle j|\,K_k^\dagger\right)\,, \tag{108}$$

$$= \sum_k\left(\sum_i|i\rangle\otimes K_k\,|i\rangle\right)\left(\sum_j\langle j|\otimes\langle j|\,K_k^\dagger\right)\,, \tag{109}$$

$$\mathcal{C} = \sum_k\mathbf{vec}(K_k)\mathbf{vec}(K_k)^\dagger\,, \tag{110}$$

where **vec** is a linear map between $d\times d$ matrices (e.g., K_k) and bipartite states (e.g., $|\Phi\rangle$) defined by

$$\mathbf{vec}(|i\rangle\langle j|) = |j\rangle\otimes|i\rangle\,. \tag{111}$$

This is an explicit form of the Choi-Jamiołkowski isomorphism. It is important to note that this "vectorization" is distinct from the vectorization introduced in Sec. II C 2. The **vec** operation associates a matrix acting on Hilbert space $\mathcal{H}$ with a vector in $\mathcal{H}\otimes\mathcal{H}$ (e.g., $|j\rangle\otimes|i\rangle$), whereas the vectorization in Sec. II C 2 merely constructs a concrete representation of that matrix as an element of $\mathcal{B}(\mathcal{H})$ (e.g., $||i\rangle\langle j|\rangle\rangle$). If the bipartite state in Eq. (111) is represented concretely, then it is possible to choose a particular basis for which these two concrete representations coincide. For example, in the computational basis,

$$\mathbf{vec}\begin{pmatrix}a & b\\ c & d\end{pmatrix} \doteq \begin{pmatrix}a\\ c\\ b\\ d\end{pmatrix}\,. \tag{112}$$

D. Models of quantum measurements

If a quantum system could not be observed, its state would be meaningless. Observations of quantum systems are called *measurements*. Sections II A 2 and II B 2 introduced models for *terminating* measurements (PVMs and POVMs) that can be used when the quantum system is used up during the measurement (e.g., photodetection) or can be thrown away (e.g., readout that concludes a quantum computation). But if the measured quantum system persists and might be observed again postmeasurement, the POVM formalism is not sufficient to predict both the measurement outcome *and* the postmeasurement state. In the context of quantum computing (and thus QCVV), such measurements are usually called *mid-circuit measurement (MCM)s*. In this section, we introduce *quantum instruments* that model MCMs, and we discuss continuous weak measurements that model the internal dynamics of the readout process.

1. Quantum instruments

Quantum measurements typically change quantum states. This is called *measurement back action* [52]. Moreover, consecutive quantum measurements can give rise to geometric phases contingent upon the order of measurements [53]. The POVM formalism, which maps a quantum state into a classical probability distribution, $\rho\mapsto\{p(i|\rho)\}$, is inadequate to describe the measurement-induced state dynamics. To address this limitation, the *quantum instrument (QI)* formalism is introduced, providing an extended framework that accounts for quantum operations influenced by the measurement outcome [54,55]. Quantum instruments can model the three-step quantum measurement procedure [56,57] introduced by John von Neumann:

(1) Initialize meter state to $|m_0\rangle$.
(2) Apply interaction between the system and meter.

(3) Read meter state.

Without loss of generality, the measurement interaction transforms the system-meter state of $\rho \otimes |m_0\rangle\langle m_0|$ into

$$\mathcal{M}(\rho \otimes |m_0\rangle\langle m_0|) = \sum_{ij} \mathcal{M}_{ij}(\rho) \otimes |m_j\rangle\langle m_j| \,, \quad (113)$$

and then the meter is projected onto one of the orthonormal eigenstates $\{|m_i\rangle\}$. The meter reads m_i with a probability $p(i|\rho)$,

$$p(i|\rho) = \mathrm{Tr}[\mathcal{M}_{ii}(\rho)] \,, \quad (114)$$

and the projection also "collapses" the principal system into the a postmeasurement state,

$$\rho_i = \frac{\mathcal{M}_{ii}(\rho)}{p(i|\rho)} \,. \quad (115)$$

A QI models this process by describing the transformation of a quantum state into a composite quantum-classical state, represented as $\rho \mapsto \{[\rho_i, p(i|\rho)]\}$. A QI $\mathcal{I}$ is a CPTP map that transforms a state ρ as:

$$\rho \mapsto \mathcal{I}(\rho) = \sum_i \mathcal{M}_i(\rho) \otimes |m_i\rangle\langle m_i| \,. \quad (116)$$

Each of the conditional quantum operations $\{\mathcal{M}_i\}$ is CP, and if there is no loss, the overall TP condition is imposed as follows:

$$\sum_i \mathrm{Tr}[\mathcal{M}_i(\rho)] = 1 \,. \quad (117)$$

2. Quantum nondemolition measurements

A *quantum nondemolition (QND)* measurement is a quantum measurement that extracts information about a system while disturbing its quantum state as little as possible. QND measurements still "collapse" quantum states into a specific eigenspace of the observable that was measured, but performing repeated QND measurements will consistently produce identical outcomes, and will not change the expectation value of the measured observable [58,59]. The reproducible and minimally perturbing nature of QND measurements plays a crucial role in achieving high fidelity readout [60,61] and is integral to quantum computing protocols including syndrome measurements in quantum error correction [62], qubit recycling [63], and algorithms for quantum machine learning [64].

The QND property can be expressed in terms of the quantum instrument (QI) formalism [Eq. (116)]. Consistency of outcomes across repeated QND measurements implies that the measurement probabilities satisfy

$$p(i|\rho_i) = \mathrm{Tr}[\mathcal{M}_i(\rho_i)] = \mathrm{Tr}[\mathcal{M}_i(\mathcal{M}_i(\rho_i))] = 1 \,. \quad (118)$$

Therefore, a QND measurement's conditional operations must be projectors:

$$\mathcal{M}_i^2 = \mathcal{M}_i \Longrightarrow \mathcal{M}_i(\rho) = \pi_i \rho \pi_i, \quad (119)$$

where the π_i are mutually orthogonal projectors. If the system experiences a Hamiltonian, then the post-measurement states must remain unchanged under the state evolution governed by the system Hamiltonian H_s,

$$e^{-iH_s t/\hbar} \pi_i e^{iH_s t/\hbar} = \pi_i \,. \quad (120)$$

This holds true if and only if, for all i,

$$[\pi_i, H_s] = 0 \,. \quad (121)$$

Interestingly, the condition given by Eq. (120) can be eased if the system exhibits periodicity, such that $e^{-iH_s\tau/\hbar} = e^{i\phi}\mathbb{I}$, where τ is the measurement interval. Therefore, even when $[\pi_i, H_s] \neq 0$, it becomes feasible to conduct a QND measurement by measuring the system at intervals of τ. This approach is known as a stroboscopic QND measurement [65].

3. Continuous weak measurements

Consider a dispersive interaction [66] between a qubit and a cavity described by the following approximated Hamiltonian:

$$H_{\mathrm{disp}} = \hbar\chi\sigma_z \otimes a^\dagger a \,, \quad (122)$$

where χ is the dispersive shift in frequency of the qubit, and $a^\dagger$ and a are creation and annihilation operators for the cavity mode. We take the initial qubit state to be $|\psi\rangle$ and, for simplicity, we assume that the cavity initially contains a coherent meter state $|\alpha\rangle$ with no energy loss (more realistic models and analyses can be found in Ref. [67]). After the qubit-meter state $|\psi\rangle|\alpha\rangle$ evolves under the dispersive Hamiltonian for time t, they become entangled as

$$e^{-iH_{\mathrm{disp}}t/\hbar}|\psi\rangle|\alpha\rangle = \Pi_0|\psi\rangle|e^{-i\alpha\chi t}\alpha\rangle + \Pi_1|\psi\rangle|e^{i\alpha\chi t}\alpha\rangle \,, \quad (123)$$

where $\Pi_i = |i\rangle\langle i|$ are the projectors onto the computational basis states. If the interaction time t is sufficiently long and the amplitude α is large enough to satisfy

$|\langle e^{i\alpha\chi t}\alpha|e^{-i\alpha\chi t}\alpha\rangle|^2 = 0$, the measurement is QND because the measurement operators are projectors that commute with the qubit Hamiltonian, since $H_{\text{disp}} \propto \sigma_z$.

In practical measurements, however, the continuous readout of the meter state introduces uncertainties due to quantum and classical sources of noise. After a short interaction time, if $|\langle e^{i\alpha\chi t}\alpha|e^{-i\alpha\chi t}\alpha\rangle|^2 \neq 0$, the meter readout of $|\beta\rangle$ becomes uncertain, resulting in a nonprojective measurement described by $\mathcal{M}(\rho) = K_\beta \rho K_\beta^\dagger$, where

$$K_\beta = \langle\beta|e^{-i\alpha\chi t}\alpha\rangle\Pi_0 + \langle\beta|e^{i\alpha\chi t}\alpha\rangle\Pi_1 \,. \tag{124}$$

Such quantum measurements are commonly referred to as *weak* measurements because, while providing some information about the system, they do not completely collapse the system to its eigenstates at once [68]. Nevertheless, successive weak measurements consistently alter the system according to $\cdots K_{\beta_2} K_{\beta_1} \rho K_{\beta_1}^\dagger K_{\beta_2}^\dagger \cdots$ and guide the state toward a specific eigenstate. Intriguingly, the state trajectory can be deduced from the measurement outcomes $\{\beta_n\}$ through the quantum Bayesian approach [69,70]. From the perspective of characterizing quantum computers, continuous weak measurements can be employed to monitor system dynamics [71,72], perform quantum process tomography [73], and diagnose gate errors [74]. The trade-off between information gain, state disturbance, and the reversibility of weak measurements has been extensively studied in Refs. [75,76].

E. Gate set models of quantum computers

Gate-based quantum computers are devices that implement quantum circuits, which are sequences of instructions for applying logic operations to physical qubits. These instructions generally include a discrete set of quantum gates, as well as state preparation, and terminating (and possibly intermediate) measurements. In the preceding sections, we have described mathematical models of all of these operations. For many QCVV protocols, it is convenient to construct a single mathematical object called a *gate set* that contains representations of all of the native instructions for a quantum device.

Formally, a gate set is the union of three distinct sets. The first one lists the N_ρ possible initial states that can be natively prepared, $\left\{\rho^{(i)}\right\}_{i=1}^{N_\rho}$. Often, quantum computers only provide a single initialization state (e.g., $|0\rangle\langle 0|$), in which case $N_\rho = 1$. The second set is a list of the computer's N_{G} native operations or *gates*, $\{G_i\}_{i=1}^{N_{\text{G}}}$. The third set lists the computer's N_{M} native measurement outcomes (POVM effects), $\left\{E_i^{(m)}\right\}_{m=1,i=1}^{N_{\text{M}},N_{\text{E}}^{(m)}}$, where $N_{\text{E}}^{(m)}$ is the the number of possible outcomes of the mth measurement. Many quantum computers offer a single native measurement in the computational basis of n qubits, in which case $N_{\text{M}} = 1$ and $N_{\text{E}} = 2^n$. The entire gate set is thus

$$\mathcal{G} = \left\{\left\{\rho^{(i)}\right\}_{i=1}^{N_\rho}, \{G_i\}_{i=1}^{N_{\text{G}}}, \left\{E_i^{(m)}\right\}_{m=1,i=1}^{N_{\text{M}},N_{\text{E}}^{(m)}}\right\} . \tag{125}$$

A gate set describes a *specific, limited* set of operations. A quantum processor may be capable of implementing other operations that are not listed in a particular gate set. A gate set $\mathcal{G}$ can only be used to describe and predict circuits built from the operations in $\mathcal{G}$.

In the context of gate sets, the word "gate" indicates an operation acting on the entire computer. The existing gate set formalism is not consistent with the alternative meaning of "gate" to denote an operation acting only on a subsystem (e.g., one or two qubits) of a quantum computer, which can be combined in parallel (by tensor product) with other gates on disjoint subsystems to produce a whole-computer operation called a circuit *layer* or *cycle*. In the gate set formalism, each layer (configuration of parallel gates) that can be performed constitutes a distinct "gate." Gate sets do not generally assume any connection or correlation between the actions of, for example, an X gate on qubit 1, an X gate on qubit 2, and parallel X gates on qubits 1 and 2. Treating each layer as an independent operation makes it possible—by comparing and contrasting different parallel combinations of gates—to study the effects of crosstalk on a device [77,78].

A gate set can be expressed in any representation that is convenient for the task at hand, but the most common convention is to use the transfer matrix representation and Hilbert-Schmidt space notation introduced in Sec. II C 2 to represent initialization operations as superkets $|\rho\rangle\!\rangle$, logic gates G as transfer matrices, and POVM measurements as lists of effects $\{\langle\!\langle E_i|\}$. Using these representations, a general gate set is written as

$$\mathcal{G} = \left\{\left\{|\rho^{(i)}\rangle\!\rangle\right\}_{i=1}^{N_\rho}, \{G_i\}_{i=1}^{N_{\text{G}}}, \left\{\langle\!\langle E_i^{(m)}|\right\}_{m=1,i=1}^{N_{\text{M}},N_{\text{E}}^{(m)}}\right\} . \tag{126}$$

In Sec. II E 1, we discuss how this representation can be conveniently used to predict quantum circuit outcomes.

A *gate set model* is a gate set-valued function of some parameters—i.e., a map from a list of parameters to gate sets. A *fully parameterized* gate set model assigns a free parameter to each matrix element of each operation in a gate set. For an n-qubit processor, a fully parameterized gate set model contains $4^n - 1$ parameters per initial state, $16^n - 4^n$ parameters per logic gate, and $8^n - 4^n$ parameters per projective measurement.

Reduced models can be constructed that use fewer parameters [78–80], motivated by sparse matrix ansätze, structural properties of the processor's Hilbert space [45], or knowledge of its low-level physics. These models have been proposed as a way to overcome the exponential

growth of parameters with system size, and their construction is an area of active research. They have fewer parameters, but generally rely on assumptions, such as limited or no crosstalk, symmetries, or *ad hoc* ansätze like low-rank tensor networks. Physics-informed reduced models can have the additional advantage of more easily interpretable parameters, such as the intensity, frequency, or phase of a control field.

Finding the parameters of a gate set model (whether fully parameterized or reduced) that fit and describe data from a particular device is the task of *gate set tomography*, discussed in detail in Sec. VII D.

1. Circuits

A gate set is a model of a quantum computer that can be used to predict the measurement outcome distribution for arbitrary quantum circuits composed of elements of the gate set. For a circuit that comprises (i) preparing native state ρ, (ii) applying the sequence of operations $C = (g_1, g_2, \ldots, g_L)$, and (iii) measuring the POVM $\{E_i\}$, the probability of measurement outcome i is given by Born's rule as

$$p(i|\rho, C) = \langle\langle E_i | G_{g_L} G_{g_{L-1}} \cdots G_{g_1} | \rho \rangle\rangle . \tag{127}$$

This can be written in the more familiar form [see Eq. (56)],

$$p(i|\rho, C) = \mathrm{Tr}[E_i \mathcal{E}_C(\rho)] , \tag{128}$$

where $\mathcal{E}_C$ is the quantum operation defined by the sequence of gates C.

2. Gauge ambiguity

A gate set model is a complete description of a Markovian quantum processor, but it is actually an *over-complete* description. A gate set contains extra *gauge* degrees of freedom that have no effect at all on *any* observable probabilities, and therefore cannot be observed. No experiment can reveal information about a gauge parameter. A *gauge transformation* on a gate set changes the gate set without changing any observable property.

A gauge transformation can be described by an arbitrary invertible $d^2 \times d^2$ matrix M, and transforms the gate set as follows:

$$\langle\langle E_i^{(m)}| \mapsto \langle\langle E_i^{(m)}| M^{-1} , \tag{129}$$

$$|\rho\rangle\rangle \mapsto M|\rho\rangle\rangle , \tag{130}$$

$$G_i \mapsto M G_i M^{-1} . \tag{131}$$

This transformation maps the gate set $\mathcal{G}$ to a new gate set $\mathcal{G}'$ with a new set of parameters, but $\mathcal{G}$ and $\mathcal{G}'$ predict identical outcome probabilities for all possible circuits because

$$\begin{aligned} &\langle\langle E | M^{-1} M G_{g_L} M^{-1} \cdots M^{-1} M G_{g_1} M^{-1} M | \rho \rangle\rangle \\ &\quad = \langle\langle E | G_{g_L} \cdots G_{g_1} | \rho \rangle\rangle \end{aligned} \tag{132}$$

for all pairs of state preparations and measurement outcomes $\{\rho, E\}$, and all sequences of gates $g_1, \ldots, g_L$. Gauge freedom implies the existence of equivalence classes of gate set models (a.k.a. *gauge orbits*) that are physically indistinguishable. As an example, Appendix E provides a introduction to gauge ambiguities in Pauli noise learning (Sec. IX C).

Gauge degrees of freedom within gate set models can significantly complicate comparisons between two models, because popular gate-error metrics like diamond norm and fidelity are gauge-dependent (see Sec. IV). One approach to mitigate these metrics' gauge dependence is to employ gauge fixing. This is most commonly done via "gauge optimization," which varies over all possible gauge transformations to find a gauge that minimizes the deviation between a (noisy) gate set model and an ideal "target" model. The metric of deviation is somewhat arbitrary, but weighted Frobenius distance is commonly used for convenience. The need for gauge-fixing can be avoided by using strictly gauge-invariant metrics of error. Gauge transformations do not change the eigenvalues of a gate's transfer matrix, so any error metric that depends only on the transfer matrix's spectrum is gauge-invariant.

Some work has explored alternative model constructions that circumvent the gauge problem. For example, Ref. [81] employs a representation of gate sets in terms of the probabilities of linear inversion gate set tomography (see Sec. VII D). This parameterization is overcomplete, and somewhat inconvenient, but completely avoids gauge freedom because every parameter in the representation is explicitly gauge-invariant. Other work [46,82] makes use of "first-order gauge-invariant" (FOGI) parameterizations that are invariant under small gauge transformations. This is an active area of research.

III. COMMON ERRORS IN QUANTUM COMPUTERS

Markovian errors in quantum computing can be broadly placed into two categories: *coherent errors* and *incoherent noise*. Coherent errors describe a reversible (purity-preserving) process in which an imperfect or unwanted unitary operator rotates the quantum register to the wrong state relative to the intended target state. Coherent errors can manifest from imperfections in gate calibrations, classical crosstalk signals that unintentionally drive a qubit, or unwanted coupling between qubits. Incoherent noise, on the other hand, describes irreversible processes, which are often referred to as *decoherence*.

The design and analysis of quantum devices must account for various intrinsic noise sources that can lead to different types of errors within the systems [83–85]. In this section, we introduce the following commonly encountered errors and noise, and illustrate their impact on single-qubit states using the Bloch sphere:

(a) *Coherent errors* (Sec. III A). When a unitary operation (including the idle) is implemented incorrectly but reversibly, the quantum register experiences a unitary (a.k.a. *coherent* or *Hamiltonian*) error. In the case of a single qubit, the qubit's state will be rotated to an incorrect point on the Bloch sphere. Coherent errors preserve purity, and can be caused by control miscalibration or entangling Hamiltonians between neighboring qubits that produce crosstalk.

(b) *Dephasing noise* (Sec. III B). Qubits in a superposition state can experience noise which leads to phase decoherence over time. For example, fluctuations in qubit frequency cause the qubit's Bloch vector to precess randomly with respect to the rotating frame, leading to random phase errors. This results in the *dephasing* of superposition states, which is visualized as shrinking of the Bloch vector towards the polar axis of the Bloch sphere.

(c) *Amplitude damping* (Sec. III C). A qubit in an excited state will eventually thermalize to its ground state. This energy relaxation process is dictated by the underlying physics of the qubit—i.e., whether it is an atom, superconducting qubit, spin qubit, etc.—but is often termed "spontaneous emission," as this is the physical pathway by which qubits thermalize for many systems. Therefore, the amplitude (or probability) of remaining in the excited state is *damped* over time. Amplitude damping is an example of a *nonunital* error, because it does not preserve the maximally mixed state.

(d) *Depolarizing noise* (Sec. III D). When incoherent noise acts isotropically about the Bloch sphere (i.e., all states have an equal probability of experiencing bit- and phase-flip errors), a qubit will eventually undergo decoherence, resulting in the complete loss of quantum information. This process is called *depolarizing* noise, because it results in the depolarization of the Bloch vector toward the center of the Bloch sphere.

(e) *Stochastic Pauli Noise* (Sec. III E). Noise in many systems is biased such that random bit- or phase-flips about different axes can occur with different rates. Such noise can be modeled by random (or stochastic) Pauli errors, whereby each type of Pauli error (e.g., X, Y, or Z) has a distinct probability of occurring.

(f) *Leakage* (Sec. III F). Qubits are defined by two computational basis states (see Sec. II A 3). *Leakage* describes any process by which a qubit transitions out of the computational subspace, either via random noise, or some coherent driving process. Leakage is often considered a *non-Markovian* process in the context of qubit computations because it cannot be described by a CPTP map on the computational subspace, and it can produce temporal correlations across multiple gates or clock cycles.

(g) *Non-Markovian and unmodeled errors* (Sec. III G). Sometimes an error in a quantum state, gate, or measurement cannot be captured by any CPTP model. In such cases, these *unmodeled errors* are typically ascribed to some *non-Markovian* process in the system. Non-Markovian errors are not the focus of this tutorial, but understanding their impact on Markovian error models is important in QCVV.

A. Coherent errors

Single-qubit unitary rotation operators $U_{\hat{\mathbf{n}}}(\theta)$ rotate a state vector $|\psi\rangle$ by an angle θ about an axis $\hat{\mathbf{n}}$. The resulting quantum state can be written as

$$|\psi'\rangle = U_{\hat{\mathbf{n}}}(\theta)\,|\psi\rangle = e^{-i\frac{\theta}{2}\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}}\,|\psi\rangle\,, \tag{133}$$

where $\boldsymbol{\sigma}$ is the vector of Pauli operators. Rotations about the coordinate axes of the Bloch sphere are particularly common, and their representations as unitary operators are

$$U_{\hat{\mathbf{x}}}(\theta) = R_x(\theta) \doteq \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & -i\sin\left(\frac{\theta}{2}\right) \\ -i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{134}$$

$$U_{\hat{\mathbf{y}}}(\theta) = R_y(\theta) \doteq \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & -\sin\left(\frac{\theta}{2}\right) \\ \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{135}$$

$$U_{\hat{\mathbf{z}}}(\theta) = R_z(\theta) \doteq \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}, \tag{136}$$

where we have used the fact that

$$e^{-i\frac{\theta}{2}\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}} = \cos\left(\frac{\theta}{2}\right) - i(\hat{\mathbf{n}}\cdot\boldsymbol{\sigma})\sin\left(\frac{\theta}{2}\right), \tag{137}$$

and where the notation $R_n(\theta)$ is commonly used in place of the notation $U_{\hat{\mathbf{n}}}(\theta)$.

Unitary (or coherent) *errors* manifest as unwanted or imperfect unitary rotations acting on qubits. This can be modeled as an ideal operator $U_{\hat{\mathbf{n}}}(\theta)$ followed by an erroneous operator $U_{\hat{\mathbf{m}}}(\epsilon)$, such that the actual final state is given by

$$|\psi'\rangle = U_{\hat{\mathbf{m}}}(\epsilon)U_{\hat{\mathbf{n}}}(\theta)\,|\psi\rangle\,, \tag{138}$$

$$= e^{-i\frac{\epsilon}{2}\hat{\mathbf{m}}\cdot\boldsymbol{\sigma}}e^{-i\frac{\theta}{2}\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}}\,|\psi\rangle\,, \tag{139}$$

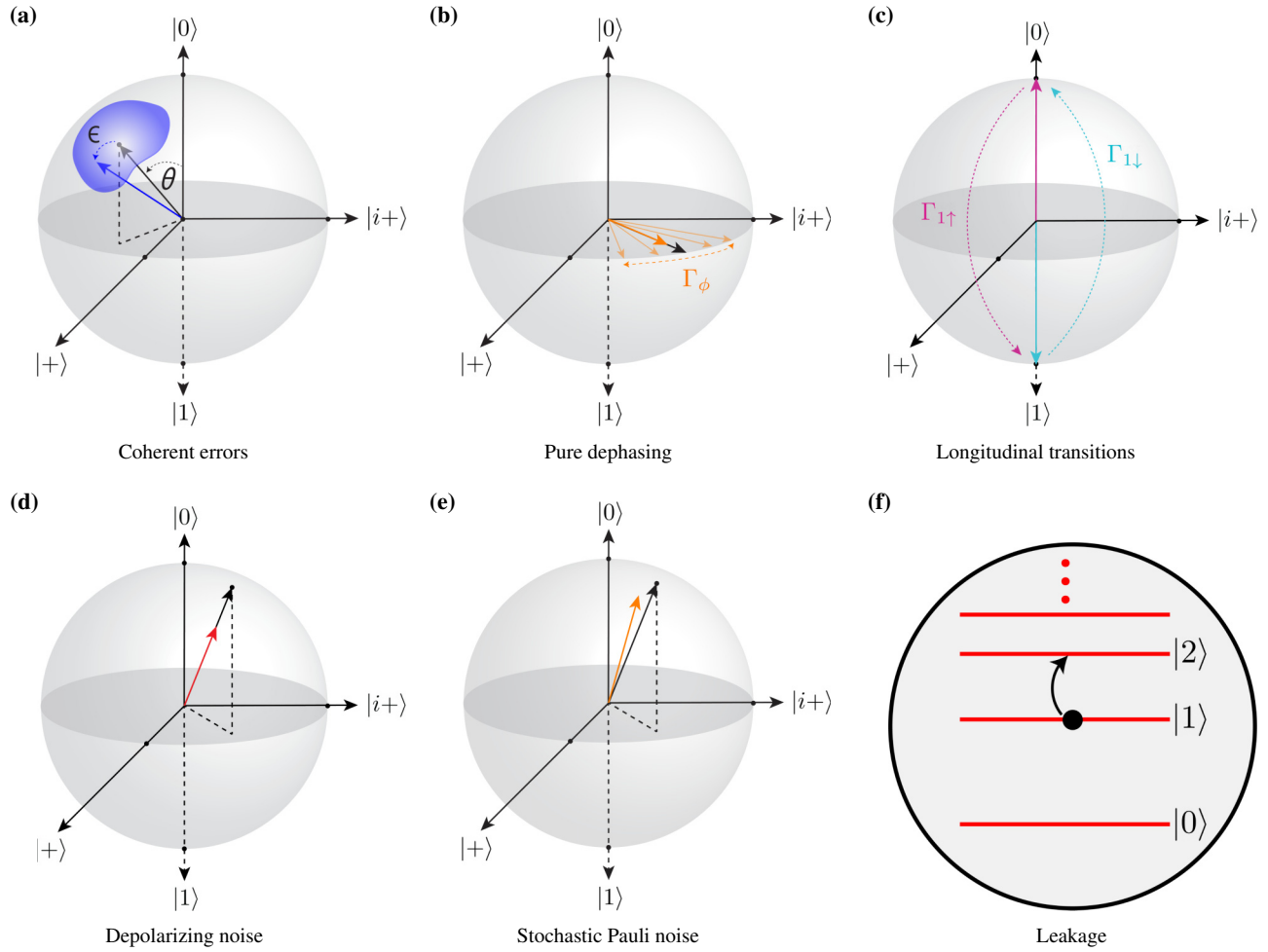


FIG. 4. Common errors and noise. (a) Coherent errors result in an unintended rotation by an angle ϵ (blue arrow) relative to the intended target state (black arrow). The axis of rotation can act in any direction relative to the intended target state, depicted by the large blue region. (b) Dephasing noise can result from fluctuations in the qubit ω_{01} transition frequency, such that the qubit Bloch vector along the equator precesses randomly with respect to the rotating frame of the qubit (light orange arrows), shrinking the Bloch vector at a rate of Γ_ϕ (orange). (c) Longitudinal transitions result from spontaneous decay (cyan) at a rate $\Gamma_{1\downarrow}$ or spontaneous excitation (purple) at a rate $\Gamma_{1\uparrow}$. (d) Depolarizing noise acts isotropically around the Bloch sphere, shrinking the length of the Bloch vector (red) relative to a pure quantum state on the surface of the Bloch sphere (black). (e) Stochastic Pauli noise acts anisotropically around the Bloch sphere, shrinking the length of the Bloch vector and resulting in an offset (orange) relative to the intended quantum state (black). (f) Leakage describes the excitation of a qubit out of the computational basis $\{|0\rangle, |1\rangle\}$ into higher energy levels.

where $\hat{\mathbf{m}}$ can be arbitrary relative to $\hat{\mathbf{n}}$. When $\hat{\mathbf{m}} = \hat{\mathbf{n}}$, as is common for certain calibration errors, the rotation axis is correct, but the rotation *angle* is wrong; this is called an over- or under-rotation error. A coherent error changes where the state vector is located on the Bloch sphere relative to the intended target state, but has no effect on the length of the Bloch vector and therefore maintains the purity of the state; see Fig. 4(a).

In the Kraus representation, a coherent error by an angle θ is given by

$$\mathcal{E}(\rho) = U_{\hat{\mathbf{m}}}(\theta)\rho U_{\hat{\mathbf{m}}}(\theta)^\dagger = e^{-i\frac{\theta}{2}\hat{\mathbf{m}}\cdot\boldsymbol{\sigma}}\rho e^{i\frac{\theta}{2}\hat{\mathbf{m}}\cdot\boldsymbol{\sigma}}, \tag{140}$$

where $K = U_{\hat{\mathbf{m}}}(\theta) = e^{-i\frac{\theta}{2}\hat{\mathbf{m}}\cdot\boldsymbol{\sigma}}$ is the Kraus operator. Below, we list various superoperator representations for a coherent error about the Z axis with Kraus operator $K = R_z(\theta)$ [see Eq. (136)]:

(a) Transfer matrix (basis of matrix units):

$$\Lambda_c = R_z(\theta)^* \otimes R_z(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{i\theta} & 0 & 0 \\ 0 & 0 & e^{-i\theta} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{141}$$

(b) PTM:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) & 0 \\ 0 & \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{142}$$

(c) χ matrix (Pauli basis):

$$\chi \doteq \frac{1}{2}\begin{pmatrix} 1+\cos(\theta) & 0 & 0 & i\sin(\theta) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -i\sin(\theta) & 0 & 0 & 1-\cos(\theta) \end{pmatrix}. \tag{143}$$

(d) Choi matrix (computational basis):

$$\mathcal{C} = \mathbf{vec}[R_z(\theta)]\mathbf{vec}[R_z(\theta)]^\dagger, \tag{144}$$

$$\doteq \begin{pmatrix} 1 & 0 & 0 & e^{-i\theta} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ e^{i\theta} & 0 & 0 & 1 \end{pmatrix}. \tag{145}$$

B. Dephasing

Dephasing is the loss of phase coherence in a quantum state. This manifests as the decay in the absolute value of the off-diagonal entries of the system's density matrix. In many physical qubit implementations, the $|0\rangle$ and $|1\rangle$ states are chosen to be energy eigenstates. The relative phase of these two states will evolve in time at a rate proportional to their energy difference. In order to perform coherent operations, external control fields must be resonant (or near resonant) with this transition, meaning that the oscillation frequency of the control fields should equal the phase evolution frequency of the qubit. When noise causes a complete loss of phase coherence between these two oscillators (the qubit and the control field), the qubit is said to have *dephased*. This can happen if either (or both) of the oscillators suffer from fluctuations in their oscillation frequency, leading to uncertainty in their relative phase [see Fig. 4(b)]. Qubits can experience frequency uncertainty due to changes in their energy splittings, such as magnetic field fluctuations, or coupling to other quantum systems, such as neighboring qubits, ac Stark shifts from control amplitude fluctuations, paramagnetic defects in semiconductors, or even the electromagnetic vacuum field. Control systems can similarly suffer a range of errors that lead to frequency instability, such as finite laser linewidths, acoustic noise in fiber optics, or clock jitter in arbitrary waveform generators.

Irreversible dephasing can arise when the qubit and clock relative frequency is changing quickly compared to the characteristic control timescale. In this case, off-diagonal entries of the density matrix are seen to decay exponentially with a characteristic timescale T_2 ("T two"; see Sec. VI D 2). T_2 is sometimes called the *transverse relaxation time* or the *intrinsic dephasing time*. In the absence of spontaneous emission effects (discussed in Sec. III C), the T_2 time is the inverse of the *pure dephasing rate*, $\Gamma_\phi = 1/T_2$.

If the relative frequency is changing slowly compared to the control timescale, then frequency errors can build up coherently for some time, and the decay of quantum coherence is nonexponential (e.g. Gaussian) rather than exponential. The characteristic timescale T_2^* ("T two star") is known as the *effective transverse relaxation time* or *inhomogeneous dephasing time*. Because the errors are correlated in time, dynamical decoupling and refocusing tools, such as the Hahn echo [86], can be used to extend the phase coherence time. The coherence decay timescale after refocusing is typically used as an estimate of the intrinsic dephasing time, and is denoted T_{2E} ("T two echo"). Protocols for characterizing the T_2^* and T_{2E} times are introduced in Sec. VI D 2. See [87–90] for background on more advanced dynamical decoupling schemes.

The Kraus representation of dephasing noise is

$$\mathcal{E}(\rho) = \left(1-\frac{p}{2}\right)\rho + \frac{p}{2}Z\rho Z, \tag{146}$$

with Kraus operators $K_I = \sqrt{1-p/2}I$ and $K_Z = \sqrt{p/2}Z$. Here, a quantum state ρ under goes a phase-flip with probability $p/2$, and is unchanged with probability $1-p/2$. Below, we list various superoperator representations for dephasing noise occurring with probability $p/2$:

(a) Transfer matrix (basis of matrix units):

$$\Lambda_c = K_I^* \otimes K_I + K_Z^* \otimes K_Z = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-p & 0 & 0 \\ 0 & 0 & 1-p & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{147}$$

(b) PTM:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-p & 0 & 0 \\ 0 & 0 & 1-p & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{148}$$

(c) χ matrix (Pauli basis):

$$\chi \doteq \begin{pmatrix} 1-p/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p/2 \end{pmatrix}. \tag{149}$$

(d) Choi matrix (computational basis):

$$\mathcal{C} = \mathbf{vec}(K_I)\mathbf{vec}(K_I)^\dagger + \mathbf{vec}(K_Z)\mathbf{vec}(K_Z)^\dagger , \quad (150)$$

$$\doteq \begin{pmatrix} 1 & 0 & 0 & 1-p \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1-p & 0 & 0 & 1 \end{pmatrix} . \quad (151)$$

C. Amplitude damping (spontaneous emission)

Many physical qubit species use two nondegenerate energy eigenstates to store quantum information, with the $|1\rangle$ state often higher in energy than the $|0\rangle$ state. Because of this energy gap, the qubit can experience *spontaneous emission*, a process in which an excited system decays to a lower-energy state by the emission of a photon, or similar (but non-radiative) energy-loss processes. These effects lead to a loss of quantum information, and are typically modeled as an *amplitude damping* error. Amplitude damping can also model the reverse process, where a qubit absorbs energy from the environment. Together, the combination of the emission and decay processes describe *thermalization*.

Strong amplitude damping on a qubit maps all points on (and within) the Bloch sphere to a single pure state, making amplitude damping the paradigmatic example of a *nonunital* process—it does not preserve the maximally mixed state. Weaker (partial) amplitude, amplitude damping errors are characterized by their decay rate. If amplitude damping describes a decay from $|1\rangle$ to $|0\rangle$, we denote the decay rate $\Gamma_{1\downarrow}$. The rate of the reverse process is denoted $\Gamma_{1\uparrow}$ (see Fig. 4). The characteristic thermalization rate, also frequently called the *longitudinal* relaxation rate, is their sum:

$$\Gamma_1 = \frac{1}{T_1} = \Gamma_{1\uparrow} + \Gamma_{1\downarrow} . \quad (152)$$

Here T_1, the "T-one time," is the characteristic thermalization timescale. If the temperature of the environment is small relative to the qubit energy splitting—i.e., $k_B T \ll \hbar\omega_{01}$—then the decay rate will be significantly larger than the absorption rate, $\Gamma_{1\downarrow} \gg \Gamma_{1\uparrow}$, and $T_1 \simeq 1/\Gamma_{1\downarrow}$. Energy decay will also impact the phase coherence of the qubit, since a qubit that decays to the ground state loses all information about its prior phase. A well-known bound [91] on the dephasing timescale is

$$T_2 \leq 2T_1 . \quad (153)$$

In Sec. VI D, we introduce protocols for characterizing both timescales.

The energy decay rate can be derived from low-level physics models through the use of Fermi's golden rule:

$$\Gamma_{1\downarrow} = \frac{1}{\hbar^2} |\langle 0|\hat{\mathcal{C}}|1\rangle|^2 \mathcal{S}_\alpha(\omega_{01}) , \quad (154)$$

where $\hat{\mathcal{C}}$ is the coupling operator to an environmental bath α at the qubit frequency ω_{01}, which is described by the noise spectral density $\mathcal{S}_\alpha$. Careful engineering of the noise environment—e.g., by placing qubits in cavities—has been shown to significantly extend T_1 times in superconducting qubits [18,92–95]. Optical-frequency qubits in atomic systems have T_1 times typically on the order of seconds, while hyperfine atomic qubits can have radiative T_1 times approaching the age of the universe. This long lifetime is due to a combination of weak magnetic dipole coupling to the electromagnetic field, and the relatively small density of states available to photons at low (microwave) splittings.

The Kraus representation of spontaneous emission (amplitude damping) is

$$\mathcal{E}(\rho) = K_0 \rho K_0^\dagger + K_1 \rho K_1^\dagger , \quad (155)$$

with Kraus operators

$$K_0 = \sqrt{I - p\sigma_+\sigma_-} = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-p} \end{pmatrix}$$

and

$$K_1 = \sqrt{p}\sigma_- = \begin{pmatrix} 0 & \sqrt{p} \\ 0 & 0 \end{pmatrix} ,$$

where $\sigma_+ = |1\rangle\langle 0|$ and $\sigma_- = |0\rangle\langle 1|$. Each Kraus operator represents an event that can occur, transforming the state. K_1 can only occur if the system is initially in $|1\rangle$, and if it occurs then a quantum of energy is lost to the environment, leaving the system in $|0\rangle$. Otherwise K_0 occurs, and the amplitude for the system to be in $|1\rangle$ gets smaller. Below, we list various superoperator representations for an amplitude damping process with decay probability p:

(a) Transfer matrix (basis of matrix units):

$$\Lambda_c = K_0^* \otimes K_0 + K_1^* \otimes K_1 ,$$

$$= \begin{pmatrix} 1 & 0 & 0 & p \\ 0 & \sqrt{1-p} & 0 & 0 \\ 0 & 0 & \sqrt{1-p} & 0 \\ 0 & 0 & 0 & 1-p \end{pmatrix} . \quad (156)$$

(b) PTM:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{1-p} & 0 & 0 \\ 0 & 0 & \sqrt{1-p} & 0 \\ p & 0 & 0 & 1-p \end{pmatrix} . \quad (157)$$

(c) χ matrix (Pauli basis):

$$\chi \doteq \frac{1}{4}\begin{pmatrix} \left(1+\sqrt{1-p}\right)^2 & 0 & 0 & p \\ 0 & p & -ip & 0 \\ 0 & ip & p & 0 \\ p & 0 & 0 & \left(1-\sqrt{1-p}\right)^2 \end{pmatrix}. \tag{158}$$

(d) Choi matrix (computational basis):

$$\mathcal{C} = \mathbf{vec}(K_0)\mathbf{vec}(K_0)^\dagger + \mathbf{vec}(K_1)\mathbf{vec}(K_1)^\dagger, \tag{159}$$

$$\doteq \begin{pmatrix} 1 & 0 & 0 & \sqrt{1-p} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & p & 0 \\ \sqrt{1-p} & 0 & 0 & 1-p \end{pmatrix}. \tag{160}$$

In the PTM representation, we can directly observe that amplitude damping is a nonunital process, because the first column is not $[1,0,0,0]^{\mathrm{T}}$.

D. Depolarizing noise

Depolarizing noise describes the process in which a quantum state ρ is replaced by a completely mixed state with some probability p,

$$\begin{aligned}\mathcal{E}(\rho) &= p\mathbb{I}/d + (1-p)\rho \\ &= \left(1-\frac{3p}{4}\right)\rho + \frac{p}{4}(X\rho X + Y\rho Y + Z\rho Z).\end{aligned} \tag{161}$$

Its the Kraus operators are $K_I = \sqrt{1-3p/4}I$, $K_X = \sqrt{p/4}X$, $K_Y = \sqrt{p/4}Y$, and $K_Z = \sqrt{p/4}Z$. Depolarizing noise acts isotropically on the Bloch sphere, and Pauli X, Y, and Z errors have an equal probability of occurring for all states. Therefore, depolarizing noise scales the length of the Bloch vector by the depolarizing probability p; see Fig. 4(d). Note that depolarizing noise is often written in a more intuitive way,

$$\mathcal{E}(\rho) = \left(1-p'\right)\rho + \frac{p'}{3}(X\rho X + Y\rho Y + Z\rho Z), \tag{162}$$

where we take $p' = 3p/4$. In this form, we may interpret a depolarizing noise channel as one in which the qubit experiences an X, Y, and Z error, each with probability $p'/3$, but remains unchanged with probability $1-p'$. Below, we list various superoperator representations for depolarizing noise:

(a) Transfer matrix (basis of matrix units):

$$\begin{aligned}\Lambda_c &= \left(1-\frac{3p}{4}\right) I\otimes I \\ &\quad + \frac{p}{4}\left(X\otimes X + Y^*\otimes Y + Z\otimes Z\right), \\ &= \begin{pmatrix} 1-p/2 & 0 & 0 & p/2 \\ 0 & 1-p & 0 & 0 \\ 0 & 0 & 1-p & 0 \\ p/2 & 0 & 0 & 1-p/2 \end{pmatrix}.\end{aligned} \tag{163}$$

(b) PTM:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-p & 0 & 0 \\ 0 & 0 & 1-p & 0 \\ 0 & 0 & 0 & 1-p \end{pmatrix}. \tag{164}$$

(c) χ matrix (Pauli basis):

$$\chi \doteq \begin{pmatrix} 1-3p/4 & 0 & 0 & 0 \\ 0 & p/4 & 0 & 0 \\ 0 & 0 & p/4 & 0 \\ 0 & 0 & 0 & p/4 \end{pmatrix}. \tag{165}$$

(d) Choi matrix (computational basis):

$$\mathcal{C} = \sum_{P\in\{I,X,Y,Z\}} \mathbf{vec}(K_P)\mathbf{vec}(K_P)^\dagger, \tag{166}$$

$$\doteq \begin{pmatrix} 1-p/2 & 0 & 0 & 1-p \\ 0 & p/2 & 0 & 0 \\ 0 & 0 & p/2 & 0 \\ 1-p & 0 & 0 & 1-p/2 \end{pmatrix}. \tag{167}$$

E. Stochastic Pauli noise

Stochastic Pauli noise generalizes both depolarizing and dephasing noise, allowing all three of the Pauli X, Y, and Z errors to have distinct probabilities p_X, p_Y, and p_Z, respectively. The Kraus representation of stochastic Pauli noise is

$$\begin{aligned}\mathcal{E}(\rho) &= (1-p_X-p_Y-p_Z)\rho + p_X X\rho X + p_Y Y\rho Y \\ &\quad + p_Z Z\rho Z,\end{aligned} \tag{168}$$

with Kraus operators $K_I = \sqrt{1-p_X-p_Y-p_Z}I$, $K_X = \sqrt{p_X}X$, $K_Y = \sqrt{p_Y}Y$, and $K_Z = \sqrt{p_Z}Z$, subject to p_X, p_Y, $p_Z \geq 0$ and $p_X + p_Y + p_Z \leq 1$. Dephasing noise is the special case where $p_X = p_Y = 0$, and depolarizing noise is the special case where $p_X = p_Y = p_Z$.

Stochastic Pauli noise is unital [i.e., $\mathcal{E}(\mathbb{I}) = \mathbb{I}$], and for a single qubit it shrinks the Bloch vector anisotropically. This reduces the Bloch vector's length, but because the

shrinking is anisotropic it can also change the Bloch vector's *direction* relative to the intended target state in a way that depends on the relative probabilities of the Pauli errors and the location of the vector on the Bloch sphere [see Fig. 4(e)]. The various representations of stochastic Pauli noise with probabilities p_X, p_Y, and p_Z are given by the following, where $p' = p_X + p_Y + p_Z''$ and $p_I = 1 - p'0$:

(a) Transfer matrix (basis of matrix units):

$$\Lambda_c = p_I I \otimes I + p_X X \otimes X + p_Y Y^* \otimes Y + p_Z Z \otimes Z = \begin{pmatrix} p_I + p_Z & 0 & 0 & p_X + p_Y \\ 0 & p_I - p_Z & p_X - p_Y & 0 \\ 0 & p_X - p_Y & p_I - p_Z & 0 \\ p_X + p_Y & 0 & 0 & p_I + p_Z \end{pmatrix}. \tag{169}$$

(b) PTM:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2(p_Y+p_Z) & 0 & 0 \\ 0 & 0 & 1-2(p_X+p_Z) & 0 \\ 0 & 0 & 0 & 1-2(p_X+p_Y) \end{pmatrix}. \tag{170}$$

(c) χ matrix (Pauli basis):

$$\chi \doteq \begin{pmatrix} 1-p' & 0 & 0 & 0 \\ 0 & p_X & 0 & 0 \\ 0 & 0 & p_Y & 0 \\ 0 & 0 & 0 & p_Z \end{pmatrix}. \tag{171}$$

(d) Choi matrix (computational basis):

$$\mathcal{C} = \sum_{P \in \{I,X,Y,Z\}} \mathbf{vec}(K_P)\mathbf{vec}(K_P)^\dagger, \tag{172}$$

$$\doteq \begin{pmatrix} p_I + p_Z & 0 & 0 & p_I - p_Z \\ 0 & p_X + p_Y & p_X - p_Y & 0 \\ 0 & p_X - p_Y & p_X + p_Y & 0 \\ p_I - p_Z & 0 & 0 & p_I + p_Z \end{pmatrix}. \tag{173}$$

The χ matrix of any Pauli stochastic noise process is diagonal in the Pauli basis, and its diagonal elements can be determined directly from the probability coefficients in the Kraus representation. A Pauli channel's PTM is *also* diagonal in the Pauli basis, and so its eigenoperators are the Pauli operators [i.e., $\mathcal{E}(P) = \lambda_P P$ for each Pauli P]. The diagonal elements of the PTM for a stochastic Pauli noise process are thus called *Pauli eigenvalues*.

Equation (170) makes apparent an important connection between the Kraus and PTM representations for stochastic Pauli noise. Namely, a Pauli error Q in the Kraus representation that occurs with probability p_Q will attenuate the Pauli eigenvalue corresponding to any Pauli operator P that anticommutes with Q (i.e., $\{P, Q\} = 0$). This can be generalized in the following way: for a set of Pauli-Kraus operators $\{Q\}$ with $\sum_Q p_Q = 1$, the resulting Pauli eigenvalues $\lambda_P = \Lambda_{PP}$ in the PTM representation are given as

$$\Lambda_{PP} = 1 - 2 \sum_{Q \text{ s.t. } \{P,Q\}=0} p_Q. \tag{174}$$

The connection between Pauli errors in the Kraus representation and Pauli eigenvalues in the PTM representation will be important when we discuss QCVV protocols for Pauli noise learning (Sec. IX C).

F. Leakage

Leakage refers to any process in which a qubit or quantum register is excited out of its computational basis states (e.g., $\{|0\rangle, |1\rangle\}$) to an orthogonal state. When a qubit is encoded into the lowest energy levels of a system, the *leakage states* are higher energy levels (e.g., $|2\rangle$); see Fig. 4(f). Leakage can be coherent (preserving phases between the computational basis and the leakage state[s]), if a qubit is unintentionally driven at its $|1\rangle \to |2\rangle$ resonant frequency, or incoherent (no phase coherence is preserved between the computational space and the leakage state[s]), if the excitation is due to thermal noise.

Some authors describe leakage as a *non-trace-preserving (TP)* process, i.e., as a qubit process that maps $\rho \mapsto \rho'$ where $\mathrm{Tr}(\rho') < \mathrm{Tr}(\rho)$ for a qubit state ρ. However, the probability of observing *some* outcome will always be 1, which makes the representation of leakage as non-TP problematic. In a system that can detect leakage events (e.g., many superconducting systems; see Fig. 36)

every experiment will either return $|0\rangle$, $|1\rangle$, or $|2\rangle$. Conversely, in systems unable or not designed to detect leakage (e.g., atomic qubits measured via resonance fluorescence, where $|0\rangle$ is measured by a dark state and $|1\rangle$ is measured by a bright state), leakage events will be *misclassified* as either $|0\rangle$ or $|1\rangle$. However, in both cases, every measurement yields *some* outcome. An example of a processes which is truly non-TP is *postselection*, in which some outcomes are *discarded* after measurement.

There is no TP Kraus representation of leakage within the qubit subspace. But leakage can be modeled rigorously by including additional states, promoting a qubit to a "qudit" with a $d > 2$-dimensional Hilbert space. Leakage can then be modeled by CPTP maps acting on $d \times d$ density matrices, by introducing Kraus operators $K = \sqrt{p_{ij}}\,|j\rangle\langle i|$ that map computational states (e.g., $\{|0\rangle, |1\rangle\}$) to leakage states (e.g., $|2\rangle$). Leakage events are often subject to selection rules between the ith and jth energy level; for example, while a transition from $|1\rangle \rightarrow |2\rangle$ is responsible for leakage out of the computational basis states, a direct transition from $|0\rangle \rightarrow |2\rangle$ might be quantum mechanically forbidden, depending on the underlying physics of the system. Transitions from the leakage states back into the computational basis states are called *seepage* [100], and can also be modeled using qudit Kraus operators.

G. Non-Markovian and unmodeled errors

So far, we have only considered *Markovian* errors. In the context of gate-based quantum computing, in particular for QCVV, an error is Markovian if it can be modeled by a CPTP map (i.e., a transfer matrix or process matrix). If an operation g on n qubits is modeled by an n-qubit CPTP map G, then G can describe any Markovian errors in the implementation of g. But there are commonly encountered phenomena that G cannot model. For example, if the n "active" qubits are accompanied by one or more "spectator" qubits that G does not describe, then G cannot account for coupling between active and spectator qubits. If applying g causes a persistent effect that induces errors during subsequent operations, that also cannot be captured by a CPTP map G. Errors that violate the assumption of temporal locality—including coupling to spectator degrees of freedom—are by definition *non*-Markovian (see Sec. VII D 3). Therefore, for n active qubits, non-Markovian errors are deviations from ideal behavior that cannot be modeled by a context-independent n-qubit transfer or process matrix.

Common types of non-Markovian errors in the NISQ era include fluctuation or drift of qubit parameters (e.g., qubit transition frequency) [96] and leakage outside of the computational basis states [97–102] [see Fig. 4(f)] with a memory longer than the timescale of the gate, correlated errors [103,104] or unwanted entanglement with qubits outside of the defined n-qubit system (e.g., static ZZ coupling in superconducting qubits [105–107]), coupling to other external fluctuators (e.g., nonequilibrium quasiparticles) [108–110], qubit heating [111], and $1/f$ noise [112, 113]; see Fig. 5. However, if we were to instead enlarge our Hilbert space to include higher energy levels and more (perhaps nonlocal) qubits, then processes like leakage and unwanted entanglement are no longer non-Markovian. Therefore, in general, quantum non-Markovianity is highly dependent upon the definition of one's system and the timescales under consideration; i.e., whether observed dynamics are Markovian or not depends on how big a model the observer uses to describe them.

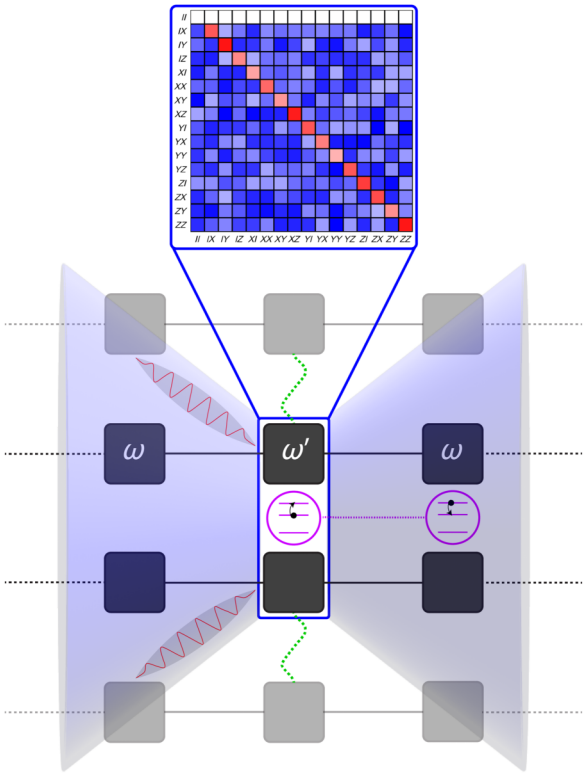


FIG. 5. Non-Markovian errors in gate-based quantum computing. For a system of two active qubits (black), all Markovian errors that occur within the timescale of a given cycle of gates (blue rectangle) can be modeled by a two-qubit transfer matrix (PTM at top of the figure; in this schematic, colored cells indicate errors in the target operation). Examples of non-Markovian errors that cannot be modeled by a two-qubit transfer matrix include (but are not limited to) drift in qubit properties over the timescale of multiple layers (e.g., fluctuations in the qubit frequency ω, with $\omega' = \omega + \delta\omega$), leakage to higher energy levels with a memory longer than the duration of a gate (purple), unwanted entanglement (green) with outside qubits (gray; e.g., due to static ZZ coupling), and classical EM crosstalk signals (red) originating from other qubits outside of the defined system that arrive within the light cone (light blue) of the system qubits. (Figure adapted with permission from Ref. [79].)

The study of quantum non-Markovianity is an active area of research [114–123], and defining *quantum non-Markovian processes* is the subject of much debate [124–128]. However, there are efforts to unify all non-Markovian processes under a common theoretical framework [129]. In this tutorial, we focus mainly on the characterization and benchmarking of *Markovian* errors, and only mention non-Markovian errors in passing. However, understanding non-Markovian errors is important for many reasons, including the fact that they are an unavoidable consequence of open quantum systems, in which the system under study is in contact with an external bath or environment with uncontrolled degrees of freedom. Additionally, non-Markovian errors interfere with the characterization of Markovian errors, which is the central goal of QCVV. Finally, their impact on quantum error correction is not well understood, which is important for fault-tolerant quantum computation.

IV. FIDELITIES AND ERROR METRICS

The purpose of QCVV is to discover and describe what is happening inside a quantum computer. In almost all cases, the computer is *intended* to do a particular thing. We call this a *target*. The models or descriptions for the target, and the thing that actually happened, can be complex and unwieldy. QCVV results are therefore often summarized by a single *performance metric* that compares what *actually* happened to what was *intended* to happen. Many such metrics exist. In this section, we introduce and explain the most common ones.

The metrics we consider in this section compare two mathematical models, one of which (the target) describes the ideal operation of the quantum computer. So, a metric [130] is generally a function $f(x,\bar{x})$, where x describes actual (i.e., experimental) behavior, and $\bar{x}$ is the target. The nature of x and $\bar{x}$ depend on what aspect of the quantum computer's operation is being examined. In this tutorial (and generally in QCVV), we consider metrics for five aspects of a quantum computer's behavior:

(1) *Probability distributions* (Sec. IV A). Probability distributions describe the results of running quantum circuits or experiments.
(2) *Quantum states* (Sec. IV B). Quantum states describe the configuration of a quantum register before it is measured.
(3) *Quantum processes* (Sec. IV C). Quantum processes describe how an operation or logic gate transforms states.
(4) *Quantum measurements* (Sec. IV D). Quantum measurements describe readout operations that extract classical data from quantum states.
(5) *Quantum processors (gate sets)* (Sec. IV E). Quantum gate sets describe a complete set of logic operations on a quantum register.

We discuss multiple distinct metrics that are commonly used for each kind of quantum object. These distinct metrics are not redundant; they quantify different aspects of an error, and each uniquely solves a particular problem. Understanding the differences between these metrics, and when to use each, is essential to reading and communicating QCVV results.

The metrics used in QCVV emerged organically. Most were borrowed or adapted from other fields of quantum information science, where their original purpose was *not* to quantify "error," but to quantify the difficulty of distinguishing two objects. As a result, their organization is somewhat haphazard. Some quantify the *similarity* of two objects. These are usually called "fidelity," and take the value $f(x,\bar{x}) = 1$ when $x = \bar{x}$. Others quantify the *deviation* between (or *distinguishability* of) two objects. These take the value $f(x,\bar{x}) = 0$ when $x = \bar{x}$, and some (but not all!) satisfy the mathematical definition of a metric. If $f(x,\bar{x})$ is a fidelity metric, then generally the corresponding *infidelity* $1 - f(x,\bar{x})$ can be used as a deviation metric. Deviation metrics, including infidelities, are often interpreted as "error rates," but we caution that no single "error rate" coincides with the probability of quantum computer failures in all contexts (which is why multiple metrics exist!).

A. Classical probability distributions

QCVV is primarily about modeling the behavior and performance of *quantum* states and processes, but they cannot be observed directly. A quantum state gains tangible reality by being measured. Quantum processes act on states, which can then be measured to yield data. Thus, the necessary common denominator in *any* experiment that tests a quantum state or process is the *probability distribution* of a measurement's outcome. For this reason, a good metric for quantum states or quantum processes is almost always derived from a more elementary metric on probability distributions. We therefore begin our survey with metrics that compare two probability distributions. However, these metrics also appear in QCVV *in their own right*, when they are used directly to evaluate the execution accuracy of a large quantum circuit.

If an experiment has a deterministic (nonrandom) outcome, and is intended to produce an outcome X, then it is very easy to test whether the experiment is working correctly. Perform a single trial, record the outcome Y, and ask whether $Y = X$. But the outcomes of quantum experiments are generally not deterministic. Both the target outcome and the actual outcome are *random* variables, described by probability distributions $\bar{\mathbf{p}}$ and $\mathbf{p}$ over some

sample space. Testing whether such an experiment is working correctly is trickier. Even in the simplest case, where the target outcome is deterministic ($\overline{\mathbf{p}}$ is supported on a single unique outcome X), we must extend the Boolean measure of correctness ("it works correctly" or "it does not") to a real-valued probability $p(X) \in [0,1]$ describing *how often* the experiment works correctly. For general $\overline{\mathbf{p}}$ and $\mathbf{p}$, quantifying the experiment's correctness becomes nontrivial. The following metrics are frequently used for this purpose.

1. Total variation distance

The *total variation distance (TVD)* between $\overline{\mathbf{p}}$ and $\mathbf{p}$ is

$$d_{\mathrm{TV}}(\overline{\mathbf{p}}, \mathbf{p}) \equiv \frac{1}{2} \sum_k |\overline{p}_k - p_k| \,, \tag{175}$$

$$= \frac{1}{2} \|\overline{\mathbf{p}} - \mathbf{p}\|_1 \,. \tag{176}$$

It has several important operational interpretations (practical questions to which it is the answer). The best-known interpretation of TVD involves *single-shot discrimination* between distributions. The optimal probability of guessing correctly whether a single sample was drawn from distribution $\overline{\mathbf{p}}$ or $\mathbf{p}$, with both theories deemed equally probable, is $[1 + d_{\mathrm{TV}}(\overline{\mathbf{p}}, \mathbf{p})]/2$. Perhaps more importantly, if N samples are drawn from distribution $\mathbf{p}$, then as $N \to \infty$ the fraction of those samples that must be *changed* in order to make them consistent with $\overline{\mathbf{p}}$ is exactly $d_{\mathrm{TV}}(\overline{\mathbf{p}}, \mathbf{p})$. This supports interpreting $d_{\mathrm{TV}}(\overline{\mathbf{p}}, \mathbf{p})$ as an *error rate*—i.e., the rate of events produced by sampling from $\mathbf{p}$ that are inconsistent with $\overline{\mathbf{p}}$.

The TVD between any two distributions is bounded between 0 (achieved uniquely when they are equal) and 1 (achieved when the distributions have disjoint support). TVD is a metric in the strict mathematical sense (e.g., it satisfies the triangle inequality).

2. Classical (Hellinger) fidelity

TVD does not capture everything. For example, consider two different discrimination problems over the set $\{0,1\}$:

(1) Distinguish $\overline{\mathbf{p}} = (1,0)$ from $\mathbf{p} = (0.98, 0.02)$,
(2) Distinguish $\overline{\mathbf{q}} = (0.49, 51)$ from $\mathbf{q} = (0.51, 0.49)$.

Both pairs are separated by the same TVD (0.02), and so can be distinguished with equal probability (0.51) given a single sample. But this is barely better than random guessing. Distinguishing either pair with reasonable confidence requires examining $N \gg 1$ samples. Rather than asking "What's the probability of guessing correctly given *one* sample?", we should ask "How many samples are required to guess correctly with high probability (e.g., 90%)?"

Remarkably, the answers for the two pairs are quite different. To distinguish $\overline{\mathbf{p}}$ from $\mathbf{p}$, we guess $\mathbf{p}$ if we see any "1" outcomes whatsoever, and $\overline{\mathbf{p}}$ if we do not. It takes just 80 samples to ensure a 90% probability of guessing correctly. But to distinguish $\overline{\mathbf{q}}$ from $\mathbf{q}$, we guess $\mathbf{q}$ if we see more "0" outcomes than "1" outcomes, and $\overline{\mathbf{q}}$ otherwise. The random fluctuations in the number of "0" and "1" outcomes are much larger in this case, and a whopping 4105 samples—greater than 50× more!—are needed to ensure a 90% probability of guessing correctly. Thus, the TVD does not *regularize* well—i.e., the TVD between $\overline{\mathbf{p}}$ and $\mathbf{p}$ does not accurately predict the TVD between the N-copy distributions $\overline{\mathbf{p}}^{\otimes N}$ and $\mathbf{p}^{\otimes N}$.

A metric that *does* regularize well is the *Bhattacharya coefficient* [131] (a.k.a., "statistical overlap" [132]):

$$BC(\overline{\mathbf{p}}, \mathbf{p}) \equiv \sum_k \sqrt{\overline{p}_k p_k} \,. \tag{177}$$

In quantum information science, the square of the Bhattacharya coefficient is often called *classical fidelity* or *Hellinger fidelity*:

$$F(\overline{\mathbf{p}}, \mathbf{p}) \equiv \left(\sum_k \sqrt{\overline{p}_k p_k} \right)^2 \,. \tag{178}$$

Classical fidelity is a measure of *similarity*: $F(\overline{\mathbf{p}}, \mathbf{p}) = 1$ if and only if $\mathbf{p} = \overline{\mathbf{p}}$, and $F(\overline{\mathbf{p}}, \mathbf{p}) = 0$ if and only if they have disjoint support. Unlike the TVD, the fidelity between two distributions does predict the fidelity between N copies of the same distributions, because

$$F(\overline{\mathbf{p}}^{\otimes N}, \mathbf{p}^{\otimes N}) = F(\overline{\mathbf{p}}, \mathbf{p})^N \,. \tag{179}$$

Classical fidelity is very closely related to the *Hellinger distance*,

$$H(\overline{\mathbf{p}}, \mathbf{p}) \equiv \sqrt{1 - \sqrt{F(\overline{\mathbf{p}}, \mathbf{p})}} \,, \tag{180}$$

which is a metric in the strict mathematical sense. The Hellinger distance is related to the TVD by a bounding inequality,

$$H^2(\overline{\mathbf{p}}, \mathbf{p}) \leq d_{\mathrm{TV}}(\overline{\mathbf{p}}, \mathbf{p}) \leq \sqrt{2} H(\overline{\mathbf{p}}, \mathbf{p}) \,. \tag{181}$$

This kind of inequality can help us understand N-copy distinguishability. If we rewrite it in terms of the classical

fidelity, we get

$$1-\sqrt{F(\overline{\mathbf{p}},\mathbf{p})}\leq d_{\mathrm{TV}}(\overline{\mathbf{p}},\mathbf{p})\leq\sqrt{2}\sqrt{1-\sqrt{F(\overline{\mathbf{p}},\mathbf{p})}}\,. \quad (182)$$

There is a strictly more powerful inequality [133],

$$1-\sqrt{F(\overline{\mathbf{p}},\mathbf{p})}\leq d_{\mathrm{TV}}(\overline{\mathbf{p}},\mathbf{p})\leq\sqrt{1-F(\overline{\mathbf{p}},\mathbf{p})}\,, \quad (183)$$

and if we apply this to the N-copy distributions, we get

$$1-F(\overline{\mathbf{p}},\mathbf{p})^{N/2}\leq d_{\mathrm{TV}}\left(\overline{\mathbf{p}}^{\otimes N},\mathbf{p}^{\otimes N}\right)\leq\sqrt{1-F(\overline{\mathbf{p}},\mathbf{p})^{N}}\,. \quad (184)$$

Therefore, the TVD between $\overline{\mathbf{p}}^{\otimes N}$ and $\mathbf{p}^{\otimes N}$ will be close to 1 if and only if $F(\overline{\mathbf{p}},\mathbf{p})^{N/2}\ll 1$—which is to say, when $N\gtrsim 2/(-\ln[F(\overline{\mathbf{p}},\mathbf{p})])$. This is the inverse of the *Bhattacharya distance*,

$$d_B(\overline{\mathbf{p}},\mathbf{p})\equiv-\frac{1}{2}\ln[F(\overline{\mathbf{p}},\mathbf{p})]\,, \quad (185)$$

which quantifies the difficulty of distinguishing very similar distributions $\mathbf{p}$ and $\overline{\mathbf{p}}$ much more accurately than the TVD. For the specific examples given in the beginning of the section, $1/d_B(\overline{\mathbf{p}},\mathbf{p})\approx 99$, whereas $1/d_B(\overline{\mathbf{q}},\mathbf{q})\approx 4999$—quite close (in both cases) to the exact number of samples required to distinguish the distributions 90% of the time.

3. Relative entropy and cross-entropy

In classical information theory, the most important and commonly used metric of deviation between distributions $\mathbf{p}$ and $\overline{\mathbf{p}}$ is neither TVD nor fidelity. Rather, it is the *Kullback-Leibler (KL) divergence*, also known as *relative entropy*:

$$d_{\mathrm{KL}}(\mathbf{p}||\overline{\mathbf{p}})\equiv\sum_k p_k\log\left(\frac{p_k}{\overline{p}_k}\right). \quad (186)$$

The KL divergence quantifies *deviation* (not *similarity*). It is always non-negative, it is zero if and only if $\mathbf{p}=\overline{\mathbf{p}}$, and it is not a mathematical metric. Unlike fidelity or TVD, d_{KL} can be arbitrarily large. Furthermore, it is *asymmetric* with respect to its two arguments, and is usually stated as "the KL divergence *from* $\overline{\mathbf{p}}$ *to* $\mathbf{p}$." Its first argument (here $\mathbf{p}$) should represent truth or reality, while its second argument (here $\overline{\mathbf{p}}$) should represent a theory or model.

The KL divergence is deeply rooted in statistics and information theory, and has too many operational interpretations to list here. It often quantifies the consequences of *believing* that samples are being drawn from $\overline{\mathbf{p}}$ when they are actually being drawn from $\mathbf{p}$. For example, it describes the rate at which a gambler or investor will lose money if they use a suboptimal strategy, the extra bandwidth required to send a message using a code adapted for the wrong distribution of symbols, and the rate at which a skeptical observer will accumulate evidence against the theory $\overline{\mathbf{p}}$ when data are actually generated by $\mathbf{p}$.

In all of these usages, the KL divergence appears as the *difference* between two quantities known as the *entropy* and *cross-entropy*:

$$H(\mathbf{p})\equiv-\sum_k p_k\log(p_k)\,, \quad (187)$$

$$H(\mathbf{p},\overline{\mathbf{p}})\equiv-\sum_k p_k\log(\overline{p}_k)\,. \quad (188)$$

The entropy of $\mathbf{p}$ (usually known as *Shannon entropy* in the information theory literature [134]) quantifies the intrinsic cost of performing a task on $\mathbf{p}$, while the cross-entropy of $\overline{\mathbf{p}}$ relative to $\mathbf{p}$ quantifies the same cost using a suboptimal strategy optimized for $\overline{\mathbf{p}}$.

Each of these quantities has many uses in its own right. The cross-entropy is particularly useful in QCVV and machine learning, because it has a rigorous mathematical meaning *and* it can be estimated easily in experiments since it is strictly linear in the true distribution $\mathbf{p}$, and can thus be written as an expectation value:

$$H(\mathbf{p},\overline{\mathbf{p}})=\langle\log(\overline{\mathbf{p}})\rangle_{\mathbf{p}}\,. \quad (189)$$

If the entropy of a candidate (model) distribution $\overline{\mathbf{p}}$ is known, then an easy way to check whether the true $\mathbf{p}$ is equal (or close) to $\overline{\mathbf{p}}$ is to estimate $H(\mathbf{p},\overline{\mathbf{p}})$ by drawing some samples, estimating $\langle\log(\overline{\mathbf{p}})\rangle_{\mathbf{p}}$, and comparing it to the known $H(\overline{\mathbf{p}})$.

4. Linear cross-entropy and heavy output probability

As noted above, cross-entropy is a well-motivated metric, with important operational interpretations, that can be measured directly. However, if $\overline{\mathbf{p}}$ has very small (e.g., zero) entries, then the variance of the estimate can be very large, making it slow to converge.

When the precise properties of cross-entropy are not important, and it is only being used as a proxy for similarity of $\mathbf{p}$ to $\overline{\mathbf{p}}$, the so-called *linear cross-entropy*,

$$H_{\mathrm{lin}}(\mathbf{p},\overline{\mathbf{p}})\equiv d\sum_{k=1}^{d}p_k\overline{p}_k-1\,, \quad (190)$$

where d is the size of the sample space, can be used instead. It is less sensitive to arbitrary small deviations in the probabilities than the real cross-entropy, but estimates of it converge with fewer samples.

Heavy output probability is another metric designed for ease of measurement. Given a d-element probability distribution $\overline{\mathbf{p}}$, the "heavy" outcomes are simply the ones

whose probability is greater than the median—i.e., the $d/2$ elements to which $\overline{\mathbf{p}}$ assigns the highest probabilities [135]. The heavy output probability of a distribution $\mathbf{p}$ with respect to $\overline{\mathbf{p}}$ is simply the total probability assigned by $\mathbf{p}$ to $\overline{\mathbf{p}}$'s heavy outcomes. Heavy output probability is easily estimated by simply drawing samples from $\mathbf{p}$ and checking whether they are "heavy" for $\overline{\mathbf{p}}$.

Linear cross-entropy [136] and heavy output probability [137] are used in QCVV to test and verify distributions over enormously large sample spaces, where sampling the entire space is infeasible. Both can be estimated fairly accurately using just a few samples. However, neither has a particularly compelling interpretation. Moreover, estimating them *does* require calculating elements of the reference distribution $\overline{\mathbf{p}}$, which can be difficult for distributions produced by quantum algorithms (the classical hardness of this task is partly why we are developing quantum computers in the first place!).

5. Roles of metrics

TVD, fidelity, and relative and cross-entropy are just a few of the many metrics, divergences, deviations, and similarity measures used in the literature to quantify similarity or distinguishability of distributions. But they are the ones that appear most frequently in the context of quantum computing and QCVV. More importantly, they are the ones from which the most commonly used properties of *quantum* objects are derived. These quantities, and their quantum counterparts discussed below, are distinct, inequivalent, and generally *not* interchangeable. When a QCVV practitioner is choosing how to quantify accuracy, error, similarity, or deviation, it is important to consider the specific task at hand. In almost every circumstance, no more than one of these quantities will faithfully capture the experimental behavior of interest.

B. Quantum states

The quantum state of a qubit, qudit, or quantum register *before* it gets measured is described by a state vector $|\psi\rangle$ (Sec. II A 1) or a density matrix $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$ (Sec. II B 1). Like probability distributions, quantum states assign probabilities to events. But for a quantum system, the sample space of possible events is determined not by the nature of the system, but by how an observer interacts with (measures) it. How similar or distinguishable two quantum states are thus depends on how they are measured. Because it is always easy to find measurements that *fail* to distinguish between quantum states, metrics for comparing two quantum states are typically defined by maximizing distinguishability, or minimizing similarity, over all possible measurements.

Metrics on quantum states can be used to compare any two states ρ and σ, but in QCVV the most common use by far is to compare a "real" state ρ to an "ideal" target state $\overline{\rho}$, and thus quantify state-preparation error. Since these metrics optimize over all possible measurements, they generally define *upper bounds* on the probability of observing an error in a specific measurement, protocol, or algorithm that uses the "real" state.

1. Trace distance

The *trace distance* between two quantum states ρ and $\overline{\rho}$ is a measure of their distinguishability. It varies from 0 (if and only if $\rho = \overline{\rho}$) to 1 (when their supports are orthogonal). It is the maximum over all POVM measurements $M = \{E_m\}$ of the TVD between M's outcome distribution given ρ, and M's outcome distribution given $\overline{\rho}$. We say that the two distributions $\Pr(m|\rho)$ and $\Pr(m|\overline{\rho})$ are *induced* by the states ρ and $\overline{\rho}$. They are given by

$$\Pr(m|\rho) = \mathrm{Tr}[E_m \rho], \tag{191}$$

$$\Pr(m|\overline{\rho}) = \mathrm{Tr}[E_m \overline{\rho}], \tag{192}$$

and so the TVD between them equals

$$d_{\mathrm{TV}} = \frac{1}{2} \sum_m \left|\mathrm{Tr}\left[E_m(\rho - \overline{\rho})\right]\right|. \tag{193}$$

Helstrom [138] proved that this is maximized by a two-outcome POVM whose effects are the projectors onto the positive and negative eigenspaces of $(\rho - \overline{\rho})$, and that the trace distance between ρ and $\overline{\rho}$ is given by the *nuclear norm* of $(\rho - \overline{\rho})$,

$$d_{\mathrm{tr}}(\rho, \overline{\rho}) = \frac{1}{2}\|\rho - \overline{\rho}\|_1 = \frac{1}{2}\mathrm{Tr}|\rho - \overline{\rho}|, \tag{194}$$

where $|\rho - \overline{\rho}| = \sqrt{(\rho - \overline{\rho})^2}$ can be obtained by diagonalizing $(\rho - \overline{\rho})$ and replacing each of its eigenvalues λ_i with its absolute value $|\lambda_i|$.

Trace distance is a metric in the rigorous sense, and a measure of distinguishability. It inherits essentially all the properties of the TVD. In particular, like TVD, it gives the probability of success for single-shot discrimination between ρ and $\overline{\rho}$. If we are given a single quantum system, prepared either according to ρ or $\overline{\rho}$ with equal prior probabilities, then the maximum achievable probability of guessing correctly how it was prepared is achieved by performing Helstrom's measurement and is equal to $[1 + d_{\mathrm{tr}}(\rho, \overline{\rho})]/2$.

The trace distance is related to the Euclidean distance between ρ and $\overline{\rho}$ for the special case of single-qubit states on the Bloch sphere. To see this, we can write ρ and $\overline{\rho}$ in

terms of their respective Bloch vectors $\mathbf{r}$ and $\bar{\mathbf{r}}$,

$$\rho = \frac{1}{2}(\mathbb{I} + \mathbf{r}\cdot\boldsymbol{\sigma}), \quad \overline{\rho} = \frac{1}{2}(\mathbb{I} + \bar{\mathbf{r}}\cdot\boldsymbol{\sigma})\,. \tag{195}$$

The trace distance between ρ and $\overline{\rho}$ is

$$d_{\mathrm{tr}}(\rho,\overline{\rho}) = \frac{1}{2}\mathrm{Tr}|\rho - \overline{\rho}| = \frac{1}{4}\mathrm{Tr}|(\mathbf{r}-\bar{\mathbf{r}})\cdot\boldsymbol{\sigma}|\,. \tag{196}$$

Because the eigenvalues of $\boldsymbol{\sigma}$ are ± 1, the trace of $|(\mathbf{r}-\bar{\mathbf{r}})\cdot\boldsymbol{\sigma}| = 2|\mathbf{r}-\bar{\mathbf{r}}|$, and thus

$$d_{\mathrm{tr}}(\rho,\overline{\rho}) = \frac{1}{2}|\mathbf{r}-\bar{\mathbf{r}}|\,. \tag{197}$$

Therefore, the trace distance between two single-qubit states is exactly equal to one-half the Euclidean distance between their Bloch vectors.

2. State fidelity

The fidelity between two quantum states ρ and $\overline{\rho}$ is a measure of their similarity. It varies from 1 (if and only if $\rho = \overline{\rho}$) to 0 (when their supports are orthogonal). In the simple special case where both states are pure, so $\rho = |\psi\rangle\langle\psi|$ and $\overline{\rho} = |\phi\rangle\langle\phi|$, their fidelity is exactly equal to the *transition probability*,

$$F(|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi|) = |\langle\psi|\phi\rangle|^2\,. \tag{198}$$

If state $|\psi\rangle$ is measured in a basis containing $\langle\phi|$, then F is the probability of observing $\langle\phi|$ and thus collapsing into $|\phi\rangle$ (and vice versa). It is important to note that the state fidelity is sometimes defined in the literature as the square root of the transition probability ($F' = \sqrt{F} = |\langle\psi|\phi\rangle|$). We (like most authors) prefer the definition given above because F is an actual probability, but readers should be aware of (and alert for) both definitions in the literature [139].

If one state is pure, but the other is mixed—e.g., $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$ and $\overline{\rho} = |\phi\rangle\langle\phi|$—then there is still a well-defined transition probability from $\rho \to |\phi\rangle\langle\phi|$. Schumacher [140] was the first to define the fidelity as

$$F(\rho, |\phi\rangle\langle\phi|) = \langle\phi|\,\rho\,|\phi\rangle = \mathrm{Tr}[\rho\,|\phi\rangle\langle\phi|]\,. \tag{199}$$

This special case is very common in QCVV, where state fidelity is commonly used to quantify error when an experimentalist intended to prepare $|\phi\rangle\langle\phi|$ but prepared ρ instead.

Defining the fidelity between two *mixed* quantum states ρ and $\overline{\rho}$ is a bit trickier because there is no obvious "transition probability" to a mixed state. There are at least two independent ways to define the fidelity between two mixed states. Remarkably, they lead to exactly the same result! Uhlmann [141] and (later) Jozsa [142] sought to generalize "transition probability" to mixed states by considering *purifications* of ρ and $\overline{\rho}$ on a larger Hilbert space. If $|\psi\rangle$ and $|\phi\rangle$ are purifications of ρ and $\overline{\rho}$ (respectively), then the *maximum* value (over all possible purifications) of the transition probability $F(|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi|) = |\langle\psi|\phi\rangle|^2$ is equal to

$$F(\rho,\overline{\rho}) = \left(\mathrm{Tr}\sqrt{\sqrt{\rho}\overline{\rho}\sqrt{\rho}}\right)^2 = \left(\mathrm{Tr}\sqrt{\sqrt{\overline{\rho}}\rho\sqrt{\overline{\rho}}}\right)^2, \tag{200}$$

which is now widely accepted as the definition of fidelity between two mixed quantum states. Fuchs [132] asked a different question that is a direct analog to Helstrom's derivation of trace distance: what is the *minimum* value, over all POVM measurements M, of the classical fidelity between $\Pr(m|\rho)$ and $\Pr(m|\overline{\rho})$? The answer turns out to be identical to Josza's fidelity [Eq. (200)].

3. Infidelity

Although Eq. (200) is celebrated, it is *very* rarely necessary in QCVV. The primary use of quantum state fidelity in QCVV is to quantify and report the *error* in an experimental attempt to prepare a pure target state $|\phi\rangle\langle\phi|$. In this situation, although the experimentally prepared state ρ is mixed, the target state is pure. The far simpler formula in Eq. (199) can be used instead.

Quantifying *error* is usually better done by reporting *infidelity* instead of fidelity:

$$\epsilon_F \equiv 1 - F\,. \tag{201}$$

Like trace distance or other measures of *distinguishability*, infidelity ranges from 0 (when $\rho = \overline{\rho}$) to 1 (when they have disjoint support). If a target state $\rho = |\phi\rangle\langle\phi|$ is prepared with infidelity 0, then the probability of an error resulting from that preparation is also zero, making infidelity a good metric of error.

Many experiments in the literature report fidelity. However, the only rationale for the awkwardness of reporting $F = 0.99981 \pm 0.00007$ instead of $\epsilon_F = (1.9 \pm 0.7) \times 10^{-4}$ are habit and a vague sense that "fidelity" plays a privileged role in the quantum information literature. This is largely a historical accident. Fault tolerance thresholds are always described by error rates, and for most QCVV purposes "low error" is a more descriptive epithet than "high fidelity."

4. Contrasting trace distance and infidelity

Trace distance and infidelity are the most commonly used metrics of deviation or distinguishability for quantum states. It is worth briefly examining what makes them different, and why neither can replace the other. Both correspond directly to classical counterparts (TVD and classical

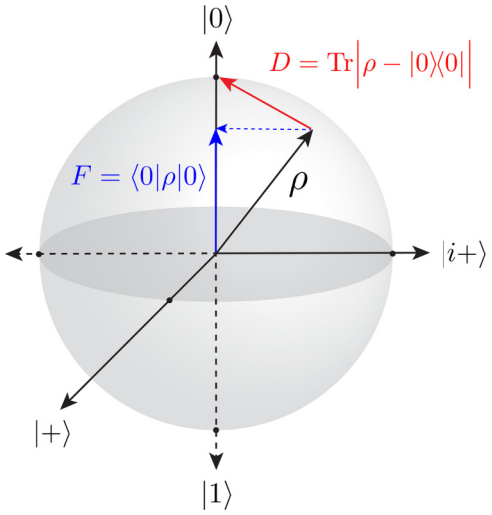


FIG. 6. Fidelity vs. trace distance. Given an arbitrary state ρ and an ideal target state $|0\rangle\langle 0|$, the "error" in ρ can be quantified in different ways. For example, the fidelity (blue) of ρ with $|0\rangle\langle 0|$ is the *projection*, or *overlap*, of ρ with $|0\rangle$: $F = \langle 0|\,\rho\,|0\rangle$. On the other hand, the trace distance between ρ and $|0\rangle\langle 0|$ is $d_{\mathrm{tr}} = \frac{1}{2}\mathrm{Tr}\,|\rho - |0\rangle\langle 0||$, which equals one-half the Euclidean distance between the two vectors on the Bloch sphere. In the figure above, D (red) is the Euclidean distance between ρ and $|0\rangle\langle 0|$, thus $D = 2d_{\mathrm{tr}} = \mathrm{Tr}\,|\rho - |0\rangle\langle 0||$. The infidelity $(1 - F)$ and the trace distance between ρ and $|0\rangle\langle 0|$ can be very different, but neither is fundamentally better at quantifying the "error" in ρ. Each is appropriate for a particular operational scenario.

fidelity) as shown by Helstrom [138] and Fuchs [132], respectively. They inherit all the properties (and differences) of those classical counterparts. So, for example, the fidelity between ρ and $\overline{\rho}$ regularizes nicely to $\rho^{\otimes N}$ and $\overline{\rho}^{\otimes N}$, whereas their trace distance does not.

But there are additional differences that appear only at the quantum level. The simplest of these have to do with the behavior of fidelity and trace distance for nearby *pure* quantum states. Suppose that $|\psi\rangle$ and $|\phi\rangle$ are "nearby" pure states, meaning that $1 - |\,\langle\psi|\phi\rangle\,|^2 = \epsilon \ll 1$. Since $|\psi\rangle$ and $|\phi\rangle$ span a two-dimensional subspace, we can consider a single-qubit system without any loss of generality (see Fig. 6). Their fidelity is $F = |\,\langle\psi|\phi\rangle\,|^2 = 1 - \epsilon$, so their infidelity is ϵ. The trace distance between them is $d_{\mathrm{tr}} = \frac{1}{2}\mathrm{Tr}\big|\,|\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\,\big|$, which we can compute by observing that because $|\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|$ has trace 0, its eigenvalues are $\{+\lambda, -\lambda\}$, and $2\lambda^2 = \mathrm{Tr}[(|\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|)^2] = 2 - 2(1 - \epsilon) = 2\epsilon$, so $\lambda = \sqrt{\epsilon}$ and thus $d_{\mathrm{tr}} = \sqrt{\epsilon}$. So, if the infidelity between two pure states is small (e.g., $\epsilon = 10^{-4}$), then the trace distance between them will be much larger ($\sqrt{\epsilon} = 10^{-2}$). It is reasonable to ask whether this behavior is generic—i.e., is it generally true that $d_{\mathrm{tr}} \approx \sqrt{\epsilon_F}$? It is not. This behavior is specific to *pure* states that differ by a unitary operation.

If instead we compare $|\psi\rangle\langle\psi|$ to a mixed state $\rho = (1 - \epsilon)\,|\psi\rangle\langle\psi| + \epsilon\,|\overline{\psi}\rangle\langle\overline{\psi}|$, where $\langle\psi|\overline{\psi}\rangle = 0$, then it is very easy to show that $d_{\mathrm{tr}} = \epsilon_F = \epsilon$. For *these* states, the two metrics coincide.

These two cases illustrate the two extremes of the *Fuchs–van de Graaf inequalities* [133], which relate infidelity and trace distance for any pair of quantum states:

$$1 - \sqrt{1 - \epsilon_F} \leq d_{\mathrm{tr}} \leq \sqrt{\epsilon_F}\,. \tag{202}$$

These inequalities are an *exact* quantum analog of the inequalities given in Eq. (183) for classical probability distributions. In the quantum case, if one state is pure, then a tighter and simpler lower bound holds:

$$\epsilon_F \leq d_{\mathrm{tr}} \leq \sqrt{\epsilon_F}\,. \tag{203}$$

5. Quantum relative entropy

Quantum relative entropy is a measure of distinguishability between quantum states. Although it is less common in the QCVV literature, it is important in quantum information theory. It generalizes KL divergence to quantum states, and is given by

$$S(\rho||\overline{\rho}) = \mathrm{Tr}[\rho\,\log(\rho)] - \mathrm{Tr}[\rho\,\log(\overline{\rho})]\,. \tag{204}$$

Like the metrics discussed above, it is defined by maximizing the *classical* KL divergence of $\Pr(m|\rho)$ with respect to $\Pr(m|\overline{\rho})$ over all POVM measurements $M = \{E_m\}$. Like the KL divergence, it is not a metric, it is asymmetric with respect to its arguments, and it diverges to infinity whenever there exists a measurement outcome such that $\Pr(m|\rho) > 0$ but $\Pr(m|\overline{\rho}) = 0$.

C. Quantum processes

Quantum computing requires that quantum states be *transformed* by precise, controlled evolution. Logic gates, circuit layers (comprising multiple gates in parallel), and quantum circuits (comprising multiple layers in sequence) describe particular *unitary* transformations that are supposed to change a d-dimensional quantum register's state as

$$\rho_{\mathrm{in}} \mapsto \rho_{\mathrm{out}} = U\rho_{\mathrm{in}}U^{\dagger} \tag{205}$$

for some $d \times d$ unitary matrix U. Real-world attempts to implement unitary transformations are imperfect, so the register's evolution is not generally described by any unitary U, but as discussed in Sec. II, it can often be described by a quantum process (CPTP map) acting on $d \times d$ density matrices,

$$\rho_{\mathrm{in}} \mapsto \rho_{\mathrm{out}} = G[\rho_{\mathrm{in}}]\,. \tag{206}$$

In this tutorial, a "quantum process" [143] means a CPTP map whose input and output spaces are the same. Such maps describe the action of gates, layers, reversible circuits, idle time, and/or imperfect unitaries. In this

section, we will refer to any such operation as a *gate*. We will denote its real (noisy) action by G, and its ideal "target" action by $\overline{G}$. The target action is almost always unitary, so $\overline{G}^{-1}$ is also a valid operation, and we can write

$$G = \mathcal{E}\overline{G} \tag{207}$$

$$\Updownarrow$$

$$\mathcal{E} = G\overline{G}^{-1}, \tag{208}$$

and refer to $\mathcal{E}$ as the *error process* for G. Most metrics for quantum processes can be used to compare two arbitrary processes, but in QCVV they are almost always used to compare a real process G to its ideal target $\overline{G}$. If a metric $f(\cdot,\cdot)$ is unitarily invariant, then $f(G,\overline{G}) = f(\mathcal{E},\mathbb{I})$.

Section II introduced several representations of quantum processes. The metrics we discuss here are properties of the quantum process itself, and their validity does not depend on what representation is being used. But each metric is most easily *defined* (and/or computed) in a particular representation. We will make extensive use of two representations (defined in Sec. II, but outlined again here):

(a) The *transfer matrix* representation Λ_G of a quantum process G (Sec. II C 2), which is constructed by choosing an orthonormal basis $\{B_i\}$ for the vector space of $d \times d$ matrices, and using the Hilbert-Schmidt inner product to define

$$(\Lambda_G)_{i,j} \equiv \mathrm{Tr}\left(B_i^\dagger G[B_j]\right). \tag{209}$$

(b) The χ *matrix* (a.k.a. "process matrix") representation χ_G of G (Sec. II C 4) is constructed by choosing an orthonormal basis $\{B_i\}$ for the vector space of $d \times d$ matrices, and then finding a matrix of coefficients $(\chi_G)_{i,j}$ such that, for any $d \times d$ density matrix ρ,

$$G[\rho] = \sum_{i,j} (\chi_G)_{i,j}\, B_i \rho B_j^\dagger. \tag{210}$$

It is easy to define *ad hoc* metrics of similarity or deviation between the matrix representations of G and $\overline{G}$. But most have no operational meaning, and are not useful. The metrics we use in QCVV and quantum computing are chosen because they have observable meanings. Since quantum processes are (like quantum states) not directly observable, meaningful metrics of similarity or deviation between G and $\overline{G}$ compare probability distributions *induced* by G and $\overline{G}$. Metrics specify (1) an initial state ρ, (2) a POVM $\{E_m\}$, and (3) a classical metric between distributions to compare

$$\Pr(m|G[\rho]) = \mathrm{Tr}(E_m G[\rho]), \tag{211}$$

$$\Pr(m|\overline{G}[\rho]) = \mathrm{Tr}(E_m \overline{G}[\rho]). \tag{212}$$

This can usually be condensed into "Choose an input state ρ and compute a known quantum state metric between $G[\rho]$ and $\overline{G}[\rho]$."

As a result, metrics for quantum processes mirror metrics for quantum states and classical distributions. The most commonly used ones are direct generalizations of TVD and classical fidelity, and inherit their properties. However, quantum processes are a richer set than states (or distributions), and display some novel behaviors. So do their metrics. In particular, *more* metrics are necessary, because there is more than one sensible way to choose a fiducial state.

1. Diamond distance

Given two processes G and $\overline{G}$, a simple natural question is "How much error would be induced by substituting G for $\overline{G}$?" There is no unique answer, because "error" is not precisely defined in this context. A more *precise* formulation is "If an *unknown* process is used just once, what is the maximum probability of guessing whether it was actually G or $\overline{G}$, given equal prior probability?" Another reasonable formulation is "If we accidentally used G in place of $\overline{G}$ in *a single spot* in a quantum information processing protocol repeated N times, what is the maximum fraction of the N outcomes that would need to be changed to cover up the mistake?"

Both framings lead to the same answer, the *diamond norm distance* (or *diamond distance*) between G and $\overline{G}$ [144]. Derived from trace distance and TVD, the diamond distance $d_\diamond(G,\overline{G})$ is the maximum trace distance between $G[\rho]$ and $\overline{G}[\rho]$, maximized over all possible input states ρ. But remarkably, the maximum value of $\left\|G[\rho] - \overline{G}[\rho]\right\|_1$ may not be attained for any *local* state ρ describing just the system on which G or $\overline{G}$ acts. A strictly higher value—and thus, greater probability of correctly distinguishing G from $\overline{G}$—can be achieved by applying the unknown process to a system that is *entangled* with another "reference" system that is not affected by the process, but can be measured jointly afterward. This counterintuitive phenomenon, akin to superdense coding [145], envariance [146], and teleportation [147], is important, because quantum logic gates are often applied to qubits that are entangled with other qubits. Restricting the maximization to local states would yield a metric that does not actually capture the worst case.

The diamond distance is defined as

$$d_\diamond(G,\overline{G}) \equiv \frac{1}{2}\left\|G - \overline{G}\right\|_\diamond, \tag{213}$$

$$\equiv \frac{1}{2}\max_{\rho_{AB}} \left\|\left((G_A - \overline{G}_A)\otimes \mathbb{I}_B\right)[\rho_{AB}]\right\|_1, \tag{214}$$

$$= \max_{\rho_{AB}} d_{\mathrm{tr}}\left((G_A \otimes \mathbb{I}_B)[\rho_{AB}], (\overline{G}_A \otimes \mathbb{I}_B)[\rho_{AB}]\right), \tag{215}$$

where A indicates the system on which G and $\overline{G}$ act, B indicates a reference system of the same dimension, and $\mathbb{I}_B$ is the identity process on the reference system. Defined this way, $d_\diamond \in [0,1]$, with $d_\diamond = 0$ if and only if $G = \overline{G}$ and $d_\diamond = 1$ if and only if they can be distinguished perfectly with a single use. In the literature, diamond distance is sometimes defined without the factor of $\frac{1}{2}$.

The diamond distance is unitarily invariant, so if $G = \mathcal{E}\overline{G}$, and $\overline{G}$ is unitary then $d_\diamond(G,\overline{G}) = d_\diamond(\mathcal{E},\mathbb{I})$. The *diamond norm error* of an error process $\mathcal{E}$ is the diamond distance between it and the identity,

$$d_\diamond(\mathcal{E}) = \frac{1}{2}\|\mathcal{E}-\mathbb{I}\|_\diamond = \frac{1}{2}\max_{\rho_{AB}} \left\| \left((\mathcal{E}_A - \mathbb{I}_A)\otimes \mathbb{I}_B\right)[\rho_{AB}]\right\|_1 . \tag{216}$$

When the "diamond norm error" of a gate is mentioned in the literature, it generally means the diamond norm error of the gate's error process (which, as noted here, is equal to the diamond distance between the gate and its target). In general, the diamond norm error of a gate is not trivial to measure, but it can be bounded by the average gate infidelity or entanglement and process infidelity of a gate (see Sec. IV C 5), which we introduce in Sec. IV C 3.

The best-known operational interpretation of $d_\diamond$ is the first one given above—it is an achievable upper bound on the probability of distinguishing G from $\overline{G}$ in a single-shot experiment. But the most *important* role of the diamond norm in quantum computing is as an error bound for circuits that use a gate *multiple* times. Aharonov *et al.* [148] showed that the diamond norm is *subadditive*. This means that if two quantum circuits (Circuit 1 and Circuit 2) are identical *except* that where operations $\overline{G}_1$ and $\overline{G}_2$ appear in Circuit 1, operations G_1 and G_2 appear instead in Circuit 2, then the diamond norm distance between the processes implemented by Circuit 1 and Circuit 2 is less than or equal to $d_\diamond(G_1,\overline{G}_1) + d_\diamond(G_2,\overline{G}_2)$.

This property, not shared by any other commonly used error metrics, makes diamond distance uniquely useful. It is often a very pessimistic upper bound on the error probability of *specific* circuits, because in many circuits (i) the initial state and final measurement are not chosen to maximize the observed TVD, and (ii) gates are arranged so that the errors in their implementation either cancel each other out (e.g., via dynamical decoupling [149,150]) or add up nonconstructively (e.g., via Pauli frame randomization [151–153] or randomized compiling [154,155]). But diamond distance provides a *guaranteed* upper bound on the accumulation of error in any quantum circuit—which can be saturated in some circumstances (e.g., error-amplifying circuits [156])—because the TVD between a circuit's ideal and experimental output distributions is bounded above by the sum of every operation's diamond norm error [144, 148]. So it is sometimes used, for example, in rigorous proofs of fault tolerance [148,157].

2. Jamiołkowski trace distance

Another metric, which is closely related to the diamond distance, is the *Jamiołkowski trace distance*. It is obtained by replacing the maximization over input states in Eq. (215) with a maximally entangled state between the system of interest and a reference of the same size:

$$d_{J-tr}(G,\overline{G}) \equiv \frac{1}{2}\left\|\left(G_A\otimes\mathbb{I}_B - \overline{G}_A\otimes\mathbb{I}_B\right)[|\Psi_{AB}\rangle\langle\Psi_{AB}|]\right\|_1 . \tag{217}$$

The Jamiołkowski trace distance provides a closed-form *lower* bound for $d_\diamond$. It is equal to the trace distance between the χ matrices of G and $\overline{G}$, and proportional to the trace distance between their Choi matrices:

$$d_{J-tr}(G,\overline{G}) = \frac{1}{2}\left\|\chi_G - \chi_{\overline{G}}\right\|_1 = \frac{1}{2d}\left\|\mathcal{C}_G - \mathcal{C}_{\overline{G}}\right\|_1 . \tag{218}$$

This relationship can be derived from Eq. (217) using the relationship between the Choi and χ representations [Eq. (100)] and the definition of the Choi representation [Eq. (97)].

3. Fidelities

The most commonly encountered performance metrics for quantum gates are *fidelities*. In fact, the word "fidelity" now transcends its technical context (like "Xerox machine" or "Kleenex"), and appears in paper titles and abstracts as a generic synonym for "quality." Despite this usage, it is still a precise technical term in quantum information science and quantum computing, and we urge readers to avoid unfortunate usage like "We quantify gate fidelity using diamond norm distance."

At least three distinct fidelity metrics appear, and are used, in the literature. The difference between them is in the initial state to which G or $\overline{G}$ is applied. But every "fidelity" quantifies *similarity*, is derived from quantum state fidelity, and inherits its properties in exactly the same way that diamond distance inherits the properties of trace distance. For every fidelity F, there is a corresponding *infidelity* $r = 1 - F$ that quantifies discrepancy and can be used as a metric of error.

a. Average gate fidelity There is a simple reason for the existence of multiple definitions of fidelity for quantum processes: quantum processes can only be "observed" by applying them to a state. The fidelity, distinguishability, or erroneousness of a process therefore depends on *context*—i.e., on what state it acts. So, the key ingredient in any definition of fidelity for quantum processes is the *output-state fidelity* for a given input state, defined in terms

of the state fidelity $F(\rho,\overline{\rho})$ [Eq. (200)] as

$$F_\rho(G,\overline{G}) \equiv F\left(G[\rho],\overline{G}[\rho]\right) . \tag{219}$$

To see why ρ matters, consider a flawed idle gate $G_{\mathbb{I}}$ that is supposed to leave states unchanged, but actually dephases them in the Z basis. If applied to a Z eigenstate ($|0\rangle$, $|1\rangle$, or any mixture of them), it acts exactly like its target, so $F_{|0\rangle\langle 0|}(G_{\mathbb{I}},\overline{G}_{\mathbb{I}}) = 1$. But if applied to an eigenstate of X or Y, it decoheres them completely, so $F_{|+\rangle\langle +|}(G_{\mathbb{I}},\overline{G}_{\mathbb{I}}) = 1/2$.

The *average gate fidelity (AGF)* eliminates this variation by the simple expedient of averaging F_ρ over all pure states using the unique (normalized) unitarily invariant Haar measure (see Appendix C 1):

$$F_{\text{avg}}(G,\overline{G}) \equiv \int F\left(G[|\psi\rangle\langle\psi|],\overline{G}[|\psi\rangle\langle\psi|]\right) \mathrm{d}\psi . \tag{220}$$

This definition applies for any G and $\overline{G}$, but $\overline{G}$ is usually unitary. If $\overline{G}[\rho] = \mathcal{U}[\rho] = U\rho U^\dagger$ for some unitary operator U, then $\overline{G}[|\psi\rangle\langle\psi|]$ is pure, and

$$F_{\text{avg}}(G,\mathcal{U}) = \int \left(\langle\psi| U^\dagger G[|\psi\rangle\langle\psi|] U|\psi\rangle\right) \mathrm{d}\psi , \tag{221}$$

$$= \int \left(\langle\langle |\psi\rangle\langle\psi|| \left(\mathcal{U}^{-1}G\right)||\psi\rangle\langle\psi|\rangle\rangle\right) \mathrm{d}\psi . \tag{222}$$

This form of the AGF makes it clear that $F_{\text{avg}}(G,\mathcal{U})$ quantifies how well the noisy process G implements the desired unitary operation U.

From the AGF, we can define the *average gate infidelity (AGI)* $r(G,\overline{G})$:

$$r(G,\overline{G}) = 1 - F_{\text{avg}}(G,\overline{G}) . \tag{223}$$

It should be noted that while $r(G,\overline{G})$ is also commonly referred to as the *average error rate* of a gate, some draw a distinction between the average error rate and average gate infidelity [158].

Many QCVV benchmarking procedures are constructed such that $U = \mathbb{I}$. In this case, the average gate fidelity is

$$F_{\text{avg}}(G) = \int \langle\psi| G[|\psi\rangle\langle\psi|] |\psi\rangle \, d\psi . \tag{224}$$

Here, F_{avg} defines the probability that G produces no detectable change in a random pure state $\rho = |\psi\rangle\langle\psi|$. Note that this is not the same as "the probability that G leaves $\rho = |\psi\rangle\langle\psi|$ unchanged," since if G deterministically rotates $|\psi\rangle \mapsto |\phi\rangle \neq |\psi\rangle$, the probability of *detecting* the change is only $1 - |\langle\psi|\phi\rangle|^2$.

The integral in Eqs. (220)–(224) can be computed explicitly [159–161] to yield a simple relationship between AGF and the (arguably more fundamental) entanglement fidelity [159,162] discussed below,

$$F_{\text{avg}}(G,\mathcal{U}) = \frac{dF_e(G,\mathcal{U}) + 1}{d+1} , \tag{225}$$

where d is the dimension of the system's Hilbert space. AGF and AGI are particularly relevant and useful in randomized benchmarking (Sec. VIII) and direct fidelity estimation (Sec. IX B), because these protocols apply a process or processes to a system initialized in randomly distributed pure—or nearly pure—*local* (unentangled) states.

b. Entanglement (process) fidelity There is another way to eliminate the state dependence of output-state fidelity. If we apply the unknown operation (G or $\overline{G}$) to a system that is maximally entangled with a reference system, so that their joint state is a maximally entangled state $|\Psi\rangle$, then the fidelity between the resulting states,

$$F_e(G,\overline{G}) \equiv F\left((G\otimes\mathbb{I})[|\Psi\rangle\langle\Psi|], (\overline{G}\otimes\mathbb{I})[|\Psi\rangle\langle\Psi|]\right) , \tag{226}$$

does not depend on *which* maximally entangled state was used. This quantity is known as the *entanglement fidelity* between G and $\overline{G}$ [163,164]. The Choi-Jamiołkowski isomorphism implies that the two states on the right-hand side of Eq. (226) are unitarily equivalent to the χ matrices χ_G and $\chi_{\overline{G}}$ (and to the trace-normalized Choi matrices $\mathcal{C}_G$ and $\mathcal{C}_{\overline{G}}$), respectively. Therefore, if $\overline{G} = \mathcal{U}$ is unitary, so that $\chi_{\mathcal{U}}$ is rank-1, then

$$F_e(G,\mathcal{U}) = \text{Tr}(\chi_G\chi_{\mathcal{U}}) = \text{Tr}\left(\chi_{\mathcal{U}^{-1}G}\chi_{\mathbb{I}}\right) . \tag{227}$$

So, the fidelity between G and a unitary target process $\overline{G} = \mathcal{U}$ equals the fidelity between the *error process* $\mathcal{E} = G\overline{G}^{-1}$ and the identity process. As discussed in the context of average gate fidelity, we often want to measure the fidelity of G with the identity operation. In this case, the entanglement fidelity is sometimes written as

$$F_e(G) = \langle\Psi| (G\otimes\mathbb{I})[|\Psi\rangle\langle\Psi|] |\Psi\rangle = \text{Tr}(\chi_G\chi_{\mathbb{I}}) . \tag{228}$$

It is sometimes [159] said that $F_e(G)$ quantifies how well G preserves entanglement, but this is not strictly correct. $\mathcal{E}$ can be entanglement-breaking, yet still have nonzero entanglement fidelity. Conversely, if G is a Pauli unitary, then $F_e = 0$ even though G does not destroy entanglement. F_e is more accurately described as the fidelity of a process *when acting on (maximally) entangled states*. For this reason, entanglement *infidelity* $(1 - F_e)$ is usually the most appropriate metric of error for quantum computing, where a gate will often act on qubits that are

TABLE II. Linear relations between performance metrics. Summary of the linear relationship between the average gate fidelity F_{avg}, the average gate infidelity r, the entanglement (process) fidelity F_e, the entanglement (process) infidelity e_F, and the process polarization f, where $d = 2^n$ for n qubits. (Table adapted from Ref. [166].)

	F_{avg}	r	F_e	e_F	f
$F_{\text{avg}} =$	F_{avg}	$1 - r$	$\frac{dF_e + 1}{d+1}$	$1 - \frac{d}{d+1}e_F$	$\frac{(d-1)f + 1}{d}$
$r =$	$1 - F_{\text{avg}}$	r	$\frac{d}{d+1}(1 - F_e)$	$\frac{d}{d+1}e_F$	$\frac{d-1}{d}(1-f)$
$F_e =$	$\frac{(d+1)F_{\text{avg}} - 1}{d}$	$1 - \frac{d+1}{d}r$	F_e	$1 - e_F$	$\frac{(d^2-1)f + 1}{d^2}$
$e_F =$	$\frac{d+1}{d}(1 - F_{\text{avg}})$	$\frac{d+1}{d}r$	$1 - F_e$	e_F	$\frac{d^2-1}{d^2}(1-f)$
$f =$	$\frac{dF_{\text{avg}} - 1}{d-1}$	$1 - \frac{d}{d-1}r$	$\frac{d^2F_e - 1}{d^2-1}$	$1 - \frac{d^2}{d^2-1}e_F$	f

entangled with other qubits. Conversion between F_e and F_{avg} is very easy using Eq. (225) (see also Table II). Entanglement fidelity is lower (more pessimistic) than average gate fidelity, because entangled states are generically more sensitive to error than random local states.

F_e is also commonly referred to as *process fidelity*. However, this usage is not entirely reliable—sometimes "process fidelity" is used to refer to other fidelity-type metrics (e.g., average gate fidelity), or as a catch-all for *any* fidelity-like metric between quantum operations. Throughout this tutorial, we only use "process fidelity" to denote F_e, but generally use (and recommend) the term "entanglement fidelity" to minimize ambiguity.

Entanglement (process) fidelity can also be computed in the Pauli transfer matrix representation (see Sec. II C 3) *if* one of the two arguments is unitary:

$$F_e(G,\mathcal{U}) = \frac{1}{d^2}\text{Tr}[\Lambda_{G\mathcal{U}^{-1}}] = \frac{1}{d^2}\text{Tr}[\Lambda_G\Lambda_{\mathcal{U}}^{-1}]\,. \tag{229}$$

The derivation is simple, starting from Eq. (227):

$$F_e(G,\mathcal{U}) = \text{Tr}(\chi_G\chi_{\mathcal{U}}) \tag{230}$$

$$= \frac{1}{d^2}\sum_{i,j,k,l}\text{Tr}(G[|i\rangle\langle j|]\mathcal{U}[|k\rangle\langle l|]\otimes|i\rangle\langle j|\,|k\rangle\langle l|)\,, \tag{231}$$

$$= \frac{1}{d^2}\sum_{i,j}\text{Tr}(G[|i\rangle\langle j|]\mathcal{U}[|j\rangle\langle i|])\,, \tag{232}$$

$$= \frac{1}{d^2}\sum_{i,j}\text{Tr}(G[|i\rangle\langle j|](\mathcal{U}^\dagger[|i\rangle\langle j|])^\dagger)\,, \tag{233}$$

$$= \frac{1}{d^2}\text{Tr}(\Lambda_{G\mathcal{U}^{-1}}) = \frac{1}{d^2}\text{Tr}(\Lambda_G\Lambda_{U^{-1}})\,. \tag{234}$$

Equivalently, if we write $\Lambda_G = \Lambda_{\mathcal{E}}\Lambda_{\mathcal{U}}$, so that $\mathcal{E}$ is the post-gate error process of the noisy operation G, then

$$F_e(G,\mathcal{U}) = F_e(\mathcal{E},\mathbb{I}) = \frac{1}{d^2}\text{Tr}[\Lambda_{\mathcal{E}}]\,. \tag{235}$$

For the remainder of this tutorial, any reference to process fidelity is a reference to Eqs. (229) or (235).

We often write χ matrices in the Pauli basis. In this basis, the χ matrix for the identity process has only one nonzero element, which is $\chi_{\mathbb{I},\mathbb{I}} = 1$. It is common to enumerate the Pauli basis elements from $0\ldots d^2-1$, starting with $\mathbb{I}$, in which case this is written as $\chi_{0,0} = 1$. The process fidelity between an error process $\mathcal{E}$ and the identity is then given by

$$F_e(\mathcal{E},\mathbb{I}) = (\chi_{\mathcal{E}})_{\mathbb{I},\mathbb{I}} = (\chi_{\mathcal{E}})_{0,0}\,. \tag{236}$$

The process *infidelity* of a gate is simply

$$e_F = 1 - F_e\,. \tag{237}$$

It is related to average gate infidelity by a simple dimension-dependent proportionality factor [159,162] (see Table II):

$$e_F = \frac{d+1}{d}r\,. \tag{238}$$

If the error process $\mathcal{E}$ has an orthogonal Kraus decomposition in which the first Kraus operator is the identity ($K_0 \propto \mathbb{I}$), then we call the error process *stochastic*, because we can model it as a probabilistic mixture of (i) no error (K_0) occurs, or (ii) an error occurs (see, e.g., Secs. III B–III E). For stochastic error processes, the process infidelity is precisely the probability that an error occurs. There is then a simple intuition for the difference between process infidelity and AGI: *every* error is detectable if it occurs on a

maximally entangled state, but if the error occurs on a random pure state, it may go undetected (e.g., if the state is an eigenstate of the error).

As a result of this, process fidelity behaves well under composition (i.e., when two gates are combined by tensor product to describe a layer of parallel gates). From Eq. (228), and the fact that a product of two maximally entangled states is maximally entangled, it follows that

$$F_e(\mathcal{E}_1 \otimes \mathcal{E}_2) = F_e(\mathcal{E}_1)F_e(\mathcal{E}_2)\,, \tag{239}$$

and therefore that

$$e_F(\mathcal{E}_1 \otimes \mathcal{E}_2) = e_F(\mathcal{E}_1) + e_F(\mathcal{E}_2) + \mathcal{O}(\epsilon^2) \tag{240}$$

if both $e_F(\mathcal{E}_1)$ and $e_F(\mathcal{E}_2)$ are $\mathcal{O}(\epsilon)$. These relationships do not hold for the AGI, because of the dimension-dependent factor. Again, this has a useful intuitive explanation: combining subsystems by tensor product increases the overall system dimension, which reduces the probability that an error will go undetected if it occurs on a random pure state.

c. Worst-case (min) fidelity One final fidelity for processes, rarely used but deserving mention, is the *stabilized minimum fidelity* [165]:

$$F_{\text{stab}}(G, \overline{G}) \equiv \min_{|\psi\rangle_{AB}} F\left((G_A \otimes \mathbb{I}_B)[|\psi\rangle\langle\psi|], (\overline{G}_A \otimes \mathbb{I}_B)[|\psi\rangle\langle\psi|]\right). \tag{241}$$

This is a fidelity-based metric (rather than a TVD-based one), but is extremized (like diamond distance) rather than averaged over input states. As with diamond distance, the minimum could also be taken over local states. But Gilchrist *et al.* observe that the resulting metric is not *stable* with respect to adding unrelated ancillary systems [165], and recommend F_{stab} instead.

Worst-case fidelity is not commonly used in QCVV. There is, to the best of our knowledge, no *good* reason for this. In many contexts, it may be better motivated than AGF or entanglement (process) fidelity. However, it suffers from sociological factors. It requires numerical computation without a nice analytic form (making it less appealing to theorists) and is strictly lower than any other fidelity metric (making it less appealing to experimentalists).

4. Process polarization

Equation (238) shows that the process infidelity and the AGI of an error process are identical up to a constant factor. A third rescaling of this quantity has a particularly intuitive use. This is the *effective depolarizing parameter* or *process polarization* of an error channel.

Many benchmarking procedures use gates in a specific way that "twirls" their error processes (see Sec. VIII A and Appendix C), effectively replacing each noisy gate $G = \mathcal{E}\overline{G}$ with $G' = \mathcal{E}_{\text{twirled}}\overline{G}$, where

$$\mathcal{E}_{\text{twirled}} = \int u\mathcal{E}u^{-1}\mathrm{d}\mu(u)\,. \tag{242}$$

In this expression, u applies a unitary transformation, and $\mathrm{d}\mu(u)$ is the normalized Haar measure over *all* unitaries acting on the gate's target Hilbert space. The effect of twirling is to symmetrize and simplify the error process drastically. It replaces $\mathcal{E}$ with a *partial depolarizing channel* $\mathcal{E}_{\text{twirled}}$ of the form

$$\mathcal{E}_{\text{twirled}} = f\,\mathbb{I} + (1-f)\mathcal{D}\,, \tag{243}$$

where $\mathcal{D}$ is the depolarizing process that acts as $\mathcal{D}[\rho] = \text{Tr}[\rho]\mathbb{I}/d$, and $1-f$ is the probability of depolarization. So,

$$\mathcal{E}_{\text{twirled}}[\rho] = f\rho + (1-f)\frac{\mathbb{I}}{d} \tag{244}$$

for any normalized density matrix ρ. We call f the *process polarization* of $\mathcal{E}$, because it quantifies the amount of polarization in ρ that remains after applying G in a context that twirls its error process.

The process polarization f is closely related to process fidelity. Analysis of twirling (see Appendix C) shows that $\mathcal{E}_{\text{twirled}}$ and $\mathcal{E}$ have exactly the same χ_{00}, and thus the same process fidelity. It is straightforward to compute that $\chi_{00} = 1$ for the identity process $\mathbb{I}$, and $\chi_{00} = 1/d^2$ for the depolarizing process $\mathcal{D}$. It follows that $F_e(\mathcal{E}_{\text{twirled}}) = f + (1-f)/d^2$, and since $F_e(\mathcal{E}_{\text{twirled}}) = F_e(\mathcal{E})$,

$$f(\mathcal{E}) = 1 - p = \frac{d^2F_e(\mathcal{E}) - 1}{d^2 - 1}\,, \tag{245}$$

where p is the probability of depolarization. Process polarization can also be computed straightforwardly from the error channel's Pauli transfer matrix as

$$f(\mathcal{E}) = \frac{\text{Tr}[\Lambda_{\mathcal{E}}] - 1}{d^2 - 1}\,. \tag{246}$$

These relationships (and others) are summarized in Table II.

A useful property of process polarization is that the polarizations of two twirled gates applied in sequence combine by simple multiplication. If $G_1 = \mathcal{E}_1\overline{G}_1$ and $G_2 = \mathcal{E}_2\overline{G}_2$, and both gates are performed in a context that twirls

them, then

$$G_2' G_1' = \mathcal{E}_{2,\text{twirled}} \mathcal{E}_{1,\text{twirled}} \overline{G}_2 \overline{G}_1 \tag{247}$$

and

$$f(\mathcal{E}_{2,\text{twirled}} \mathcal{E}_{1,\text{twirled}}) = f(\mathcal{E}_{2,\text{twirled}}) f(\mathcal{E}_{1,\text{twirled}}) \,. \tag{248}$$

Note that this simplification does not apply for *parallel* composition, because twirling is system-dependent, and the tensor product of two (locally) twirled error channels is not (globally) twirled.

5. *Contrasting diamond distance and infidelity*

As we have seen in this section, there are many different ways to quantify the "error rate" of a quantum process. Both the AGI r [Eq. (223)] and entanglement (process) infidelity e_F [Eq. (237)] have the convenient interpretation of being *average* error rates (i.e., the rate at which an error would be observed, averaged in some way over possible input states). On the other hand, TVD-derived error metrics such as the diamond distance $d_\diamond$ [Eq. (215)] are *maximizations* over all possible POVMs and/or input states, and thus are sometimes called *worst-case* error rates.

While average error rates can be efficiently measured by Monte Carlo sampling (see, e.g., Sec. VIII), estimating extremal quantities like diamond distance is harder. For example, although tomographic reconstruction methods (see Sec. VII) can be used to estimate the diamond distance [167] by means of semidefinite programs [168,169], the cost of tomography grows exponentially with the number of qubits. However, an error channel's AGI or process infidelity provides a *bound* on its diamond norm error [158,170,171]:

$$\frac{d+1}{d} r \leq d_\diamond \leq \sqrt{d(d+1)}\sqrt{r}\,, \tag{249}$$

$$e_F \leq d_\diamond \leq d\sqrt{e_F}\,, \tag{250}$$

where d is the dimension of the Hilbert space. Tighter bounds can be obtained if one has knowledge of the *unitarity* of a gate [171], which we introduce in Sec. VIII H.

To illustrate the types of errors that saturate the bounds of the diamond norm, we consider two types of single-qubit errors: (1) a coherent (unitary) error, and (2) a stochastic error. A single-qubit coherent X error corresponds to a unitary rotation by some angle θ,

$$R_X(\theta) = \begin{pmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ i\sin(\theta/2) & \cos(\theta/2) \end{pmatrix} . \tag{251}$$

The PTM superoperator of this error is given as

$$\Lambda_\mathcal{E} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} . \tag{252}$$

Considering the difference between the identity operation and the error,

$$\mathbb{I} - \Lambda_\mathcal{E} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1-\cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & 1-\cos(\theta) \end{pmatrix} , \tag{253}$$

we observe that, for small θ, the magnitude of the diagonal elements scale as $|1-\cos(\theta)| \approx \frac{1}{2}\theta^2$, and the magnitude of the off-diagonal elements scale as $|\sin(\theta)| \approx \theta$. Because the diamond norm is the maximization over all possible input states and, via the trace distance, also a maximization over all POVMs, it is sensitive to the largest elements of $\mathbb{I} - \Lambda_\mathcal{E}$. In the case of the coherent error given above, the largest elements of $\mathbb{I} - \Lambda_\mathcal{E}$ are the off-diagonal elements, which scale as $\mathcal{O}(\theta)$, and thus $d_\diamond \sim \theta$. Now, consider *twirling* $\Lambda_\mathcal{E}$ into a stochastic Pauli channel (Sec. III E) or depolarizing channel (Sec. III D) via Pauli or Clifford twirling, respectively (see Sec. VIII A and Appendix C). In both cases, the diamond norm of $(\Lambda_\mathcal{E})_\text{twirled}$ scales as its process infidelity. In this example, we observe that the diamond norm is *at least* $d_\diamond \sim e_F \approx \theta^2$ (when $\Lambda_\mathcal{E}$ represents a stochastic error channel), and *at most* $d_\diamond \sim \sqrt{e_F} \approx \theta$ (when $\Lambda_\mathcal{E}$ represents a unitary channel). Thus, it is often said that the lower bound of the diamond norm is saturated by a purely stochastic noise channel, and the upper bound of the diamond norm is saturated by a purely unitary error channel [79,171,172]. This example is only meant to be a *heuristic*—it is not a rigorous derivation of the bounds of the diamond norm—but it does highlight where the quadratic difference between the lower and upper bounds of the diamond norm come from, and how the diamond norm can differ by orders of magnitude in the presence of stochastic noise versus coherent errors.

D. Quantum measurements

Quantum measurements, or *readout operations*, are the third essential logic operation in a quantum processor. Two distinct kinds of measurement operation appear in quantum circuits and quantum computing experiments, *terminating measurements* that mark the end of a circuit (after which the entire processor may be reinitialized, cooled, and/or recalibrated before another circuit is run), and *mid-circuit measurements* (MCMs). Mid-circuit measurements are harder to implement, because they must (1) not disrupt other qubits that are not being measured, and (2) leave the measured qubit(s) in a usable state.

Terminating measurements have been studied and analyzed much more thoroughly than mid-circuit measurements in the QCVV community. However, the analysis of error metrics for both kinds of measurement is surprisingly rare in the literature. The metrics most commonly used (readout fidelity and QND-ness) are relatively ad hoc in comparison to the systematic framework that has been developed for states and processes. We outline the most commonly used metrics below.

1. Terminating measurements

Terminating measurements are modeled by POVMs (see Sec. II B 2), and have been an important subject of QCVV since at least 1999 [173]. However, the literature on error metrics for POVMs is sparse [174–177].

Quantum computing experimentalists usually seek to perform *orthogonal rank-1 projective* measurements. For each outcome i of the measurement, there is a unique pure state $|i\rangle$ for which $p(i|i) = 1$. Some of the most common metrics are specialized for this case. One example is *readout fidelity*, defined as the average probability of observing outcome i given state $|i\rangle$, or

$$F_{\text{readout}} = \frac{1}{d}\sum_{i=0}^{d-1} p(i|i) = \frac{1}{d}\sum_{i=0}^{d-1} \text{Tr}[E_i\,|i\rangle\langle i|]\,, \tag{254}$$

where d is the number of outcomes (and Hilbert-space dimension), and E_i is the POVM effect associated with measurement outcome i [see Eq. (56)]. Readout fidelity is used extensively as folklore (without citation) in the superconducting qubit literature [178–181]. It can be defined as the average of a more fundamental quantity that we denote *effect-wise fidelity*:

$$F_i = p(i|i) = \text{Tr}[E_i\,|i\rangle\langle i|]\,. \tag{255}$$

The F_i are useful when they vary substantially over outcomes i, in which case the worst-case fidelity,

$$F_{\text{min}} = \min_i F_i\,, \tag{256}$$

is relevant. For example, the excited-state readout fidelity F_1 of many systems is often worse than the ground-state readout fidelity F_0 due to energy relaxation (e.g., T_1 decay; see Sec. III C). However, these effect-wise fidelities can be equalized through methods that twirl measurement noise [182,183].

None of these quantities is directly observable, unless the experimenter has the ability to prepare perfect initial states $|i\rangle\langle i|$. In real-world experiments, it is impossible to unambiguously distinguish errors in state preparation from errors in measurement. So, these errors are often grouped together and referred to as *state preparation and measurement (SPAM)* errors [184]. Directly measured quantities generally depend (in more or less complicated ways) on

$$F_{\text{SPAM}(i)} = p(i|\rho_i) = \text{Tr}[E_i\rho_i]\,. \tag{257}$$

Error metrics for terminating measurements that are *not* supposed to be orthogonal rank-1 projective are almost nonexistent in the literature, to the best of our knowledge. Although general nonprojective POVMs are rarely implemented on purpose, nonorthogonal rank-1 measurements (e.g., SIC POVMs [185]) and non-rank-1 projective measurements (e.g., stabilizers [186]) are important use cases to which Eqs. (254) and (255) do not necessarily apply. General metrics can be obtained by observing that a POVM is a kind of CPTP map (a "quantum-classical" or q-c channel [187]), so every process metric defined previously (entanglement fidelity, diamond distance, etc.) can be computed for POVMs. However, none are in common usage (although [176] discusses many possible fidelities). For example, the entanglement fidelity between a POVM $\{E_i\}$ and the ideal POVM $\{|i\rangle\langle i|\}$ works out to

$$F_e\left(\{E_i\},\{|i\rangle\langle i|\}\right) = \left(\sum_i \sqrt{F_i}\right)^2\,, \tag{258}$$

which is not quite the same as the ubiquitous Eq. (254) (although they agree to leading order in $1 - F_i$). TVD-based metrics, though well-motivated whenever the intended output distribution is nontrivial, have not seen widespread use.

2. Mid-circuit measurements

Mid-circuit measurements (MCMs) are modeled by quantum instruments (see Sec. II D 1), and are critical for quantum error correction. Their importance to QCVV has grown rapidly in recent years. However, the literature on error metrics for MCMs is essentially limited to Ref. [188], which focuses on the important special case of *uniform stochastic instruments* and shows that entanglement fidelity and diamond distance (defined by treating the instrument as a CPTP map) are suitable metrics.

The experimental literature on characterization of MCMs (e.g., Ref. [177]) primarily reports readout fidelity as defined in Eq. (254), and another folklore metric called *QND-ness* (see Sec. II D 2, and also Refs. [189,190]), which is defined as the average probability of getting the same measurement result twice in a row:

$$Q = \frac{1}{d}\sum_{i=0}^{d-1} p(i,i|i)\,, \tag{259}$$

$$= \frac{1}{d}\sum_{i=0}^{d-1} \frac{\text{Tr}[|i\rangle\langle i|\,\mathcal{M}_i(|i\rangle\langle i|)]}{p(i|i)}\,, \tag{260}$$

where the noisy mid-circuit measurement is given by a set of completely positive (CP) maps $\{\mathcal{M}_i\}$. Like readout fidelity, QND-ness is practical, but impossible to measure exactly without the ability to prepare perfect input states. As noted in Ref. [189], QND-ness is not a very useful metric for non-rank-1 measurements, because it only measures repeatability and has no sensitivity to whether the mid-circuit measurement disrupts other observables that it should commute with.

The difference between QND-ness and readout fidelity can be illustrated by considering qubit measurements that failed to satisfy the QND requirements as described in Sec. II D 2. For instance, measurements that inherently alter a quantum state such as charge detection [191] or resonance fluorescence [192,193] may achieve high fidelity, but leave the system in a state outside the qubit manifold. Alternatively, state errors during measurement can lead to significant non-QND-ness, but have only minimal effect on the readout fidelity. For example, T_1 decay processes during dispersive readout of the excited state of superconducting qubit can be observed and corrected for by time-resolved state discrimination (this can be accomplished using weak, continuous measurement, outlined in Sec. II D 3), but the final state of the system may have decayed to the ground state. In such cases, QND-ness can often be improved by actively reinitializing the state $|i\rangle\langle i|$ corresponding to whatever i was read out, or shortening the measurement time.

E. Quantum processors (gate sets)

So far, we have considered each logic operation independently, in isolation. This is consistent with the history of the field. However, it is internally *inconsistent*—and, more importantly, unrealistic. We needed measurements to define state fidelity, states to define measurement fidelity, and *both* of them to define process fidelity. Every logic operation is only defined (and observable) *relative* to other logic operations. This became widely recognized between 2012 and 2016 [194–196]. This relationality creates a gauge freedom that couples all of a quantum processor's logic operations and requires them to be treated as a *gate set*, rather than a set of independent operations [197] (see Sec. II E).

Many modern QCVV protocols (starting with randomized benchmarking) explicitly mix together properties of all the operations in a gate set, to produce a holistic metric that quantifies the error rate not of any single operation, but of a processor's entire gate set. We outline this below.

1. Average gate-set (in)fidelity

Perhaps the most common error metric for gate sets is *average gate-set infidelity (AGSI)* [196]. This is simply the average, over all gates in a gate set (*not* including state preparation or measurement), of the AGI [Eq. (223)].

AGSI was originally believed to correspond accurately to the error rate observed in randomized benchmarking (RB) [184], which is broadly agreed to be one useful "error rate" for a quantum processor. Gauge freedom turns out to complicate this relationship [196], but if AGSI is evaluated in the gauge that minimizes the gate-to-gate variation of the individual gates' error channels, it does in fact correspond well to the RB error rate [198].

Since 2018, randomized benchmarking protocols have proliferated (see Sec. VIII for some examples), and they do not all measure the same "error rate." Therefore, it is a good idea to read the defining paper for a particular RB protocol carefully before interpreting its result! However, most RB error rates are related, at some level, to AGSI.

2. Circuit output distributions

The other class of metrics that are commonly used to evaluate the performance of processors and gate sets are actually the *classical* metrics discussed in Sec. IV A, applied to the outcome distributions of quantum circuits. In particular, linear cross-entropy [2] and heavy output probability [137] are widely used to quantify how accurately a quantum processor has executed a circuit. Other metrics (e.g., TVD [155,199] or Hellinger fidelity [200]) are also used, but less commonly.

V. DESIGN AND IMPLEMENTATION OF QCVV EXPERIMENTS

Quantum computers implement quantum algorithms by preparing quantum states, applying quantum gates, and performing quantum measurements. Characterization and benchmarking experiments each provide insight into the types and rates of the errors that affect these operations, but can be broadly distinguished by what they measure. Characterization experiments are typically designed to fit the parameters of a *statistical model* that attempts to capture some aspect of the data generating process. These models are often *interpretable*—their parameters have physical meaning—and so may be used to identify the physical source of an error. Benchmarking experiments, on the other hand, are typically designed to assess performance as captured by some empirical measure of success or accuracy, such as the average success probability of circuits specified by a particular algorithm. These performance metrics do not generally permit reliable attribution of errors to physical sources, but benchmarking protocols usually scale more efficiently to many-qubit processors than detailed characterization protocols, and may be more indicative of application performance. The distinction between characterization and benchmarking protocols is often somewhat blurred in practice.

A QCVV *protocol* can be thought of as a recipe for the design and analysis of a characterization or benchmarking experiment. For the purposes of this tutorial, a protocol takes as input a register of qubits and a set of native quantum operations, and outputs a set of (possibly randomized) quantum circuits. A protocol also specifies a data analysis procedure for fitting a model to the experimental data (in the case of characterization) or extracting a performance metric (in the case of benchmarking). Most protocols leave additional important experiment design parameters unspecified—e.g., the order in which circuits should be run, the number of shots to be taken per circuit, and the frequency of recalibration. In this section, we outline several principles that can inform these decisions, and discuss some experimental realities that can constrain them. To ensure clarity and reproducibility, papers that report QCVV results should state clearly the specific choices made when implementing QCVV protocols. We divide our discussion in two parts:

(a) *Principles of QCVV experiment design* (Sec. V A). QCVV experiments utilize particular families of quantum circuits to probe the noisy dynamics of quantum hardware. Many of these circuit families share a common structure that helps ensure the protocol is *robust* and *informationally complete*. Running these circuits on real hardware often requires *compilation* and *scheduling* that can impact the performance and results of the protocol.

(b) *Identifying and mitigating out-of-model effects* (Sec. V B). QCVV protocols are typically designed to be accurate across a wide range of experimental conditions. However, a quantum computing system may suffer errors that were not accounted for, such as leakage or non-Markovianity, that may cause biased or nonsensical results. Models can be *extended* to include these error types, or the protocols can be designed and run in a way that *averages*, *mitigates*, or *quantifies* their effects.

A. Principles of QCVV experiment design

1. Quantum circuit families

A QCVV experiment design should enable the effective and efficient study of the specific errors under test while remaining (ideally) agnostic to other sources of error that might be present in the system. The experiment design necessarily includes a family of circuits to run, which is often specified by the QCVV protocol. QCVV circuit families may comprise a finite set of specific circuits, as in state (Sec. VII A) or process tomography (Sec. VII B), or an ensemble of random circuits to be sampled according to some prescribed measure, as in randomized benchmarking (Sec. VIII).

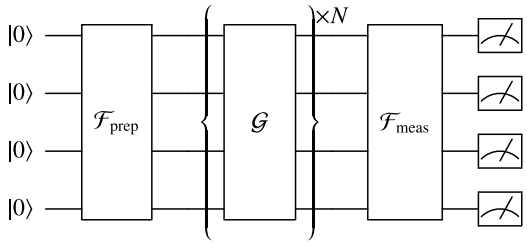


FIG. 7. Typical QCVV circuit structure. Circuits used in QCVV often comprise a short state preparation gate sequence $\mathcal{F}_{\text{prep}}$, an N-fold repeated (possibly randomized) gate sequence $\mathcal{G}$, a measurement preparation circuit $\mathcal{F}_{\text{meas}}$, and a computational basis measurement.

The circuits specified by QCVV protocols often share a common *sandwich* structure, as illustrated in Fig. 7. Such circuits typically comprise initialization in the all-zeros state, a short state-preparation gate sequence $\mathcal{F}_{\text{prep}}$ (sometimes called a *preparation fiducial circuit*), an N-fold repeated (possibly randomized) gate sequence $\mathcal{G}$, a measurement preparation circuit $\mathcal{F}_{\text{meas}}$ (sometimes called a *measurement fiducial circuit*), and a concluding computational basis measurement. This structure facilitates the estimation of specific errors: the $\mathcal{F}_{\text{prep}}$ circuit creates a state that is sensitive to some aspect of $\mathcal{G}$'s performance, the N-fold repetition of $\mathcal{G}$ amplifies some aspect of its errors, and $\mathcal{F}_{\text{meas}}$ implements a measurement that is sensitive to the target errors. Table III illustrates how this structure manifests in several common QCVV protocols.

In order to serve as effective and efficient probes of error, circuits defined by a QCVV protocol should possess a few common properties:

Informational completeness: Taken together, data from all specified circuits should be sufficient to enable the estimation of all desired parameters. *Informational completeness* occurs when there is sufficient independent data (circuit outcome statistics) to compute the protocol's performance metric or to reconstruct the target model parameters. If the circuit ensemble is *over*-complete, and the QCVV protocol reconstructs a statistical model for the data, then formal model validation can be used to assess the quality of the fit [77,156]. *Under*-complete circuit ensembles generally do not allow for the estimation of all model parameters without some additional regularization, but can be useful, for example, when doing sparse model selection, as in compressed sensing [201,202].

Formally, informational completeness of an experiment design corresponds to the associated Fisher information matrix [203,204] being full rank over the vector space of model parameters (see Sec. II E 2 for a more nuanced perspective). The Fisher information is a powerful tool for analyzing an experiment design and evaluating its ability to probe the parameters of a statistical model.

TABLE III. Examples of "sandwich" QCVV circuit structure. Many QCVV circuits possess a sandwich structure, illustrated in Fig. 7, wherein a state-preparation operation $\mathcal{F}_{\text{prep}}$ is applied first, followed by N applications of a short (possibly random) circuit $\mathcal{G}$, concluding with a measurement preparation circuit $\mathcal{F}_{\text{meas}}$ and terminating measurement of all qubits. Here, we explicitly list the components of common QCVV experiments. Unless otherwise specified, randomized gates are generally intended to be drawn uniformly at random from the Haar distribution over the associated group (see Appendix C 1) each time the randomized gate is applied within a circuit. Here, $\mathbb{C}_n$ is the set of n-qubit Clifford operations, $\mathbb{S}_n$ is the set of permutation operations on n elements, and $\mathsf{SU}(4)^{\otimes \lfloor n/2 \rfloor}$ is the $\lfloor n/2 \rfloor$-fold Cartesian product of $\mathsf{SU}(4)$, where each factor represents an arbitrary two-qubit operation acting on a pair of qubits. The $\mathsf{InverseClifford}$ operation is the measurement preparation that, conditional on a particular realization of the random Clifford $\{\mathsf{Random}(\mathbb{C}_n)\}$ elements, acts as the inverse operation, ideally returning any state to its initial value at the start of the circuit. The gate $\mathcal{G}$ for Rabi and Ramsey experiments are sometimes implemented *discretely*, as indicated in the table above, or via *continuous* driving; see Sec. VI for more details.

Experiment	$\mathcal{F}_{\text{prep}}$	$\mathcal{G}$	$\mathcal{F}_{\text{meas}}$
Rabi oscillations (Sec. VI B)	...	$X_{\pi/2}$	...
Ramsey oscillations (Sec. VI C 2)	$X_{\pi/2}$	I	$X_{\pi/2}$
Randomized benchmarking (Sec. VIII B)	...	$\mathsf{Random}(\mathbb{C}_n)$	$\mathsf{InverseClifford}$
Gate set tomography (Sec. VII D)	$\mathcal{F}_{\text{prep}}$	$\mathcal{G}$	$\mathcal{F}_{\text{meas}}$
Quantum volume (Sec. XI A 1)	...	$\mathsf{Random}(\mathbb{S}_n) \circ \mathsf{Random}\left[\mathsf{SU}(4)^{\otimes \lfloor n/2 \rfloor}\right]$	...

The Cramér-Rao bound states that the absolute precision of any estimator is bounded by the Fisher information, so circuits with a large Fisher information are therefore preferable to those with small Fisher information. Further details about the Fisher information are beyond the scope of this tutorial. The interested reader is encouraged to consult Ref. [203] for an introduction to the fundamentals of Fisher information and its applications in quantum information processing.

Amplificational completeness: Informational completeness guarantees that an experiment will have *some* sensitivity to every parameter of interest. Its sensitivity—i.e., the precision with which parameters can be estimated—can be increased by increasing the number N of experimental shots (counts). But in almost all cases, expected estimation error decreases relatively slowly, as $\mathcal{O}(1)/\sqrt{N}$, with the number of shots. A different (and often better) way to make experiments more sensitive is to grow the *length* (L) of the circuits in the experiment, "amplifying" certain parameters. Many QCVV experiments define scalable families of circuits parameterized by a nominal length L. If such an experiment's sensitivity to *all* parameters of interest grows uniformly with L, so that the expected estimation error decreases as $\mathcal{O}(1)/L$, then we say the experiment is "amplificationally complete." Amplificational completeness is usually evaluated using Fisher information [204].

Robustness: QCVV protocols are typically designed to probe particular errors or some measure of performance (e.g., fidelity; see Sec. IV). But the errors under test are rarely the only errors present in the system. QCVV protocols are often designed to suppress other phenomena while amplifying the target errors. Dynamical decoupling [149, 150], Pauli frame randomization [151–153], randomized compiling [154,155], and group twirling (see Sec. VIII A and Appendix C) are particularly well-known examples of such techniques.

Classical simulability: QCVV experiments probe the noise and errors in quantum processors by comparing observed circuit outcomes to those expected in a noiseless system, or those predicted by a noise model. This comparison often requires calculating the expected outcome distribution for a quantum circuit. The classical hardness of this problem is the entire reason we are building quantum computers! QCVV circuits use a number of techniques to preserve classical simulability. For comparing to ideal unitary evolution, these include: restriction to nonuniversal gate sets (such as Clifford circuits) [205], structured inversion [206], and restriction to small systems. When comparing to statistical noise models, additional constraints on the model are often enforced to preserve classical simulability, such as tensor-network ansätze [207] or restriction to low-weight errors [208–210]. Circuit primitives, such as group twirling, can further enhance the performance of simplified error models and improve simulability.

2. Implementation details

The precision with which an experiment can measure a parameter is controlled by the experiment design (the quantum circuits to be run), the amount of data (the number of "shots" per circuit), and the data analysis procedure (the estimator).

Among the more important experimental constraints is the time it takes to run a full QCVV experiment. Model-based characterization of multiqubit systems can be among the most experimentally taxing applications of near-term quantum computers. A single run of two-qubit gate set tomography (see Sec. VII D), for instance, can take several hours on neutral atom or trapped-ion quantum computers. During this time, the environmental degrees of freedom are likely to drift, so that data from circuits run at the beginning of data collection will reflect a different noise

environment than those from the end. Periodic gate recalibration can mitigate these effects to some extent, but the recalibration rate should be consistent with the expected drift during typical operation.

The experimentalist must also make a number of choices about the specifics of the data collection procedure that will impact the precision and reliability of the results. Some of these choices include the following:

Circuit repetitions (shots): The precision of an estimator grows with the amount of data collected, so more repetitions are often better than fewer. However, more data requires more experiment time. The experimentalist must balance the benefit of greater precision against the cost of time. Estimation error decreases only as $1/\sqrt{N}$ with the number of shots N, so more data yield diminishing returns.

Maximum circuit depth: Deeper (longer) circuits can amplify errors and therefore serve as more precise probes of gate error—particularly coherent errors—than short circuits. However, if the circuits are too long, then decoherence can reduce the visibility of the target errors. Furthermore, short circuits are useful for ensuring consistency of an estimator, and are often used to improve convergence of an optimizer in postprocessing. Several protocols, including robust phase estimation (Sec. VI E) and gate set tomography (Sec. VII D), use logarithmically spaced circuit lengths in order to strike an appropriate balance between stability and precision.

Data-collection order: Drift and/or hardware recalibration can lead to time correlations in physical error rates. If data is taken in batches (all samples of a circuit are taken in sequence before moving on to the next circuit), different circuits can experience different noise environments. This can cause bias in parameter and metric estimation and be difficult or impossible to identify *post hoc*. If data collection is instead *rastered* (data is taken in many passes, with each pass taking one shot of each circuit in the experiment design), drift effects will be smoothed across the dataset, and data can be analyzed for signs of time-correlated noise. Not all experimental systems can be configured to take rastered data, and for these an intermediate collection procedure may be necessary wherein all data is retaken in two or three batches. Such data can be used to probe for low-frequency drift [96].

Circuit compilation rules: Integrated quantum processors often feature a *compiler* that optimizes the scheduling of circuits to maximize the processor's performance. When running QCVV protocols, this can sometimes lead to confusing results. For instance, Ramsey experiments (Sec. VI C 2) use long idle periods or sequences of repeated idle gates. Care must be taken that the compiler does not identify this and remove the "extra" idles, or the expected Ramsey decay may not be observed. Similarly, it is impossible to probe certain crosstalk errors if, for example, the compiler forbids multiple two-qubit gates from acting in parallel. In some cases, programmatic "barriers" in the low-level quantum assembly code can enforce compilation restrictions. When characterizing and benchmarking quantum systems, users should take steps to ensure that the compiler is not altering quantum circuit instructions in ways detrimental to the QCVV protocol.

In addition to these considerations, experiments must contend with control system constraints, including the data-collection rate, recalibration times, memory buffer sizes, network latency, arbitrary waveform generator (AWG) upload times, constraints on batching versus rastering, and others. These constraints can prevent a QCVV protocol from collecting data optimally, and some care should be taken to consider the impact of these constraints on the reliability, susceptibility to bias, and potentially increased variance of derived performance metrics.

B. Identifying and mitigating out-of-model effects

As discussed above, characterization experiments are typically designed to learn all or some parameters of a statistical error model describing a quantum device. If physical errors are present in the experiment that are not captured by this model, then the model will not fit the observed data, and estimates of model parameters can be significantly biased by out-of-model effects. In this section, we briefly survey techniques that can be used to make QCVV protocols robust to out-of-model effects, or at least estimate their impact.

1. Extending models

The simplest approach to mitigating out-of-model effects is to modify the protocol to turn them into in-model effects. An example of this is leakage quantification (see Secs. III F and VIII I 1) using "blind" randomized benchmarking (RB) [211], a form of character benchmarking [212] (see Sec. VIII for details on randomized benchmarking and its various variants, such as character RB). Blind RB modifies the models used by standard RB to explicitly include a parameter describing the population of leakage levels that cannot be directly observed. Unobserved leakage causes the usual RB decay curve to become a mixture of *two* exponential decays rather than a single exponential decay:

$$\bar{p}(m) = A + Bf^m + Cg^m . \tag{261}$$

Fitting a mixture of exponentials can be difficult, especially in the presence of noise. Blind RB modifies the experiment design so that, rather than all circuits compiling to the identity, half of the circuits are chosen to compile to a bit-flip operation. In the absence of leakage, the success probability of these circuits should decay at the same rate as the standard identity circuits. However, the bit-flip operation

does not act on the leakage state, so the success probability decays differently:

$$\bar{p}_x(d) = A' - Bf^d + Cg^d\,. \tag{262}$$

By adding and subtracting the decay curves for the two experiments, we get two new curves that decay as single exponentials and so are easy to fit. Blind RB estimates both the error rate per Clifford and the *leakage rate* per Clifford.

Tomographic protocols (Sec. VII) can also be extended by explicitly growing the size of the model. Leakage can be captured, for instance, by modeling a qubit as a qu*trit* (or, more generally, a qu*dit*), and performing state, process, or gate set tomography on the larger model (see Appendix D 2 for some examples of tomography applied to qutrits and ququarts). This can get expensive, though, and requires careful thought if the leakage levels are not coherently addressable. This also makes it hard to construct a tomographically complete set of states and measurements, and so it may be impossible to fit a full qudit model. Instead, reduced models that assume, for example, only *incoherent* leakage may be more experimentally tractable.

2. Averaging out-of-model effects

Sometimes QCVV models are *known* to not capture all of the errors in a system. For instance, tomographic protocols often fit a static model to a system that is actually experiencing drift in some physical parameter. If the effect of drift is not mitigated by the experiment design, the tomographic estimate will be biased. For instance, drift in the measurement error rate can significantly impact the RB decay curve and subsequent error rate estimates. If data are taken in order of increasing circuit length, growing measurement errors can make the decay curve steeper than it should be, causing the error per gate to be overestimated. On the other hand, if the data is taken with circuits in *decreasing* order of length, then growing measurement errors over the course of the experiment will produce a shallower curve, and an underestimate of the error per gate.

This bias effect also manifests in tomographic routines. Bias can be reduced (at the possible cost of increased variance) by removing the correlation between execution time and circuit properties. This can be done by selecting a new circuit from the experiment list for each shot, until all repetitions and circuits have been consumed. Another approach is to *raster* the data—collecting a single shot from each circuit in some order, and then repeating until enough shots have been taken for all of the circuits. If the shots from each circuit are then averaged, the drift will be distributed uniformly across the data set. But rastered data can also be analyzed directly, using methods such as the ones introduced in Ref. [96], to yield time-dependent estimates of error rates.

3. Mitigating out-of-model effects

While the above methods can be used to average out-of-model effects, sometimes we just want to eliminate large classes of noise. Again, one particularly frustrating source of noise is drift in control parameters. This drift can lead to errors that change over the course of an experiment. In general, there are three main approaches to reducing the effects of parameter drift: recalibration, dynamical decoupling, and twirling.

When experiments are impacted by low-frequency noise (like the ubiquitous $1/f$ noise in solid state systems), periodic recalibration can dramatically reduce the scale of the drift problem. This naturally comes at the cost of experimental time, but is often necessary in long experiments to ensure that data taken over long times is consistent. Recalibration may involve fine-tuning experimental control parameters or completely rerunning the calibration procedure *ab initio*, feeding back on results from just a few carefully chosen circuits, or feeding forward data taken from spectator qubits [213,214]. What method is chosen will depend on experimental capabilities, the timescale of the experiment, and the nature of the drifting error rates.

Dynamical decoupling (DD) [149,150]—or dynamically corrected gates more generally—has a long history of eliminating the impact of unknown coherent sources of errors. DD evolved from the Hahn echo in nuclear magnetic resonance [86], where sequential radio frequency pulses are used to reverse the effects of inhomogeneities in the local magnetic field, effectively refocusing the spins and producing a detectable echo signal. In quantum computing, DD uses sequences of gate operations to cancel coherent errors or low-frequency dephasing noise (Sec. III B) whose magnitudes are unknown or drifting.

Twirling is a technique closely related to dynamical decoupling, where interleaved gate operations are used to prevent the buildup of coherent errors. Unlike DD, twirling uses random gates, and often aggregates data taken from different randomized circuit realizations (a.k.a. "randomizations"). Thus, twirling can result in a larger experimental overhead, requiring a different circuit to be measured per randomization. This can be largely mitigated by performing the twirling directly on the control hardware on a shot-by-shot basis [215]. Twirling can dramatically reduce the complexity of an error channel, and is a key component of randomized benchmarks (Sec. VIII) and methods such as randomized compiling [154,155]. Twirling is discussed in detail in Sec. VIII A and Appendix C.

4. Quantifying out-of-model effects

Even if all the techniques above are deployed, there remain scenarios in which data will display clear evidence of out-of-model effects. Models will simply not fit the data. In this circumstance, several statistical tools (e.g., likelihood ratio tests) can be deployed to detect and quantify

a model's failure to fit the data [77]. However, statistical techniques based on hypothesis testing can only determine the confidence with which a model can be rejected. They do not generally provide a measure of *effect size*, i.e., how much a model's predictions deviate from actual observations. For instance, consider a coin that is modeled as fair (i.e., 50/50), but when flipped a trillion times yields 501 billion heads and 499 billion tails. The χ^2 statistic—a simple statistical model validation tool—is approximately 4×10^6, meaning that the fair coin model can be rejected with incredibly high confidence (a p-value of practically zero [216]). But, for most practical purposes, a coin that is biased by 0.1% can be well-approximated as fair. The effect size—0.1%—is simply too small to matter for, say, a football game coin toss.

Quantum tomography experiments often provide a lot of data, so the best-fit model can often be rejected by a statistical hypothesis test. This does not necessarily mean that it is a bad model. It means that it is demonstrably not a *perfect* model. There are visible deviations from the model, which are not just statistical fluctuations, but indicate the existence of unmodeled effects. Whether the model is "good enough" should be evaluated not using a statistical confidence measure (which grows with the size of the dataset), but using some measure of the *size* of out-of-model effects.

One proposed way to quantify effect size in QCVV is to compute a wildcard error model [217]. Wildcard models have been deployed for gate set tomography (GST) experiments, where they relax GST models so that, rather than predicting a circuit's outcome probabilities, they predict a *range* of outcome probabilities. One kind of wildcard model does this by adding a bit of extra "wildcard" error to each gate. Wildcard models state that the total variation distance [Eq. (175)] between each circuit's outcome distribution and the standard GST model prediction should be no larger than the circuit's wildcard error. The per-gate wildcard error is then chosen to be minimally sufficient to make the data statistically consistent with the wildcard model predictions. The wildcard error can then be compared against the parameters of the GST model. If the wildcard is small relative to the gate error, then the GST model captures the most important noise sources, even if statistical tests indicate high confidence for rejecting the GST model. See, for example, Refs. [78,79] for how this is done in practice.

VI. QUBIT AND GATE CHARACTERIZATION

The first steps needed to run a quantum computer are to characterize basic qubit properties and calibrate quantum gates. Qubit characterization involves measuring the resonant frequency and coherence times (i.e., how long a qubit behaves quantum mechanically) of the qubit. These properties are important for calibrating the quantum gates used in algorithms. For example, the coherence times of a qubit place fundamental limits on the fidelity of quantum gates or measurements performed on that qubit, and additionally inform the end user of the circuit depth with which one can perform useful computations with that qubit. Moreover, many quantum gates require coherently driving qubits on-resonance; therefore, it is necessary to characterize qubit frequencies to high accuracy. In this section, we review standard methods for characterizing basic qubit properties and discuss how they can be utilized for measuring errors in quantum gates to high precision:

(1) *Frequency-domain spectroscopy* (Sec. VI A). The first step in probing a quantum system is to find its transition frequencies. To achieve this, one can irradiate a driving field on the qubit and sweep its frequency across a broad range. As the drive induces the qubit's transitions when it is swept across the corresponding resonant frequencies, the subsequent measurement of the qubit then reveals its energy spectrum.

(2) *Rabi oscillations* (Sec. VI B). A fundamental test of qubit control is to perform *Rabi oscillations*, whereby a qubit is coherently driven on-resonance between its ground and excited states. Measurement of the qubit after a varied drive duration reveals an oscillation pattern characteristic of a two-level quantum system.

(3) *Time-domain spectroscopy* (Sec. VI C). The coherent control of a qubit requires driving it at its resonant frequency. Therefore, accurately finding qubit frequencies is an important step in the calibration of quantum gates. We review two methods for characterizing qubit frequencies in the time domain, including *Ramsey spectroscopy*, a standard interferometric experiment that is used throughout atomic, molecular, optical, and solid state physics.

(4) *Qubit coherence* (Sec. VI D). The *coherence times* of a qubit are characterized by two timescales: (1) thermalization (or energy relaxation), which quantifies how long a qubit will remain excited before decaying to the ground state; and (2) phase relaxation, which quantifies how long a qubit in a superposition state will maintain phase coherence. Measuring these two properties can be accomplished with simple Rabi and Ramsey experiments.

(5) *Phase estimation* (Sec. VI E). Standard Rabi and Ramsey experiments are typically performed in a continuous manner (see, e.g., Fig. 9). However, in a gate-based setting, one can instead perform *discrete* Rabi and Ramsey experiments, which are constructed out of a set of defined quantum logic gates. This is the basis for a class of methods known as *phase estimation*, which can be used to perform precision measurements of small errors in the rotation angles of quantum gates.

A. Frequency-domain spectroscopy

Measuring the energy spectrum of a qubit is a foundational step in quantum characterization. Coarse measurements of qubit transition frequencies can be performed using standard laboratory spectroscopy methods, such as absorption spectroscopy [219]. When driven on-resonant using an external electromagnetic field (e.g., optical laser, microwave signal, etc.), the qubit will absorb some energy from the radiation field and undergo a transition from its initial state, resulting in a change in the measurement signal. Otherwise, the signal remains constant up to the noise level. Therefore, by sweeping the frequency of the external field and monitoring the reflected or transmitted signal, we can detect the transition frequency of the qubit. For example, Fig. 8 shows the transition spectrum from the $|0\rangle$ and $|1\rangle$ states of a superconducting transmon qubit.

If the qubit is continuously driven across a wide range of frequencies during the measurement, the technique is broadly referred to as *continuous-wave (CW) spectroscopy*. To mitigate spurious and higher-order effects from multiphoton processes, the qubit drive may be deactivated during the measurement phase, which is then termed *pulsed spectroscopy*. Together, these techniques are often called *frequency-domain spectroscopy*. While the broad linewidths found using frequency-domain spectroscopy typically provide sufficient frequency information to observe the coherent nature of qubits (see Sec. VI B), it is not generally precise enough to calibrate quantum gates; instead, one must resort to time-domain techniques, such as Ramsey spectroscopy, to obtain such information (see Sec. VI C).

While some platforms can be probed directly, such as atomic systems, where resonance fluorescence is often used to measure qubit states, other platforms are measured using an ancilla system. For example, superconducting qubits are often measured via dispersive coupling to a readout resonator [67], in which the frequency of the readout resonator is dependent on the state of the coupled qubit; thus, by probing the resonant frequency of the readout resonator, one can determine what state the qubit was in. In such cases, it is necessary to first perform frequency-domain spectroscopy on the ancilla system to characterize its resonant frequency. Then, to measure the resonant frequency of the qubit, we sweep the frequency of the field driving the qubit, while also monitoring the frequency spectrum of the readout resonator. This is referred to as *two-tone spectroscopy*, since the qubit and ancilla system often operate at different frequencies.

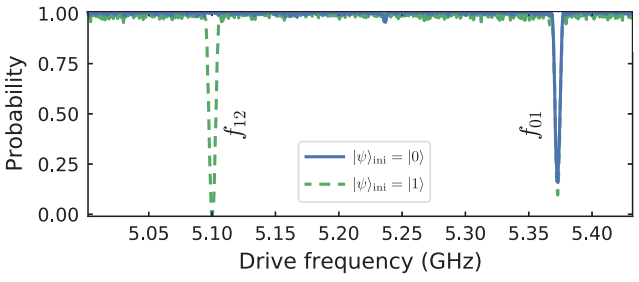


FIG. 8. Frequency-domain spectroscopy. By sweeping the frequency of a microwave tone driving a transmon qubit and monitoring the readout signal, we can detect the transitions of the qubit from its initial states, $|\psi\rangle_{\rm ini} = |0\rangle$ (solid blue line) and $|\psi\rangle_{\rm ini} = |1\rangle$ (dashed green line). The observed dips indicate the resonant frequencies, which correspond to the $|0\rangle \to |1\rangle$ and $|1\rangle \to |2\rangle$ transitions. (The data are reproduced with permission from Ref. [218].)

B. Rabi oscillations

A two-level system, such as a qubit, can be rotated about the Bloch sphere [see Fig. 1(a)] using a classical oscillating field

$$\mathbf{E} = E_0 \left(\boldsymbol{\epsilon}_d e^{-i(\omega_d t+\phi_d)} + \text{h.c.}\right), \tag{263}$$

where E_0 is the amplitude of the field, ω_d is the driving frequency, ϕ_d is the phase, $\boldsymbol{\epsilon}_d$ is the complex polarization vector, and h.c. is the Hermitian conjugate. This field is coupled to the qubit's dipole moment

$$\mathbf{D} = d\,(\boldsymbol{\epsilon}_s \sigma_+ + \text{h.c.}), \tag{264}$$

where d is the dipole matrix element, $\boldsymbol{\epsilon}_s$ is the qubit's polarization vector, and σ_+ the qubit raising operator. The coupling Hamiltonian between the qubit and the field is then given by $H_d = \mathbf{E}\cdot\mathbf{D}$. By aligning the field with the polarization of the qubit, and by performing the *rotating wave approximation (RWA)* to simplify the driving scheme, we can write the Hamiltonian in the interaction picture as

$$\begin{aligned} H_I &= \frac{\hbar\delta\omega}{2}\sigma_Z + \frac{\hbar\Omega}{2}(e^{-i\phi_d}\sigma_+ + e^{i\phi_d}\sigma_-), \\ &= \frac{\hbar\delta\omega}{2}\sigma_Z + \frac{\hbar\Omega}{2}(\cos\phi_d\sigma_X + \sin\phi_d\sigma_Y), \end{aligned} \tag{265}$$

where $\delta\omega = \omega_q - \omega_d$ is the qubit-drive detuning, and $\Omega = 2dE_0/\hbar$ is the *Rabi frequency* of the driven system, which is intuitively proportional to the dipole matrix element d and the amplitude of the field E_0.

Due to the finite detuning between the drive and qubit frequencies, the qubit does not always rotate at the Rabi frequency. More generally, we can write the interaction Hamiltonian as

$$H_I = \frac{\hbar\Omega'}{2}\boldsymbol{\sigma}\cdot\hat{\mathbf{n}}, \tag{266}$$

where $\Omega' = \sqrt{\delta\omega^2 + \Omega^2}$ is the *effective Rabi frequency*, and

$$\hat{\mathbf{n}} \equiv \frac{\delta\omega \mathbf{u}_Z + \Omega \cos\phi_d \mathbf{u}_X + \Omega \sin\phi_d \mathbf{u}_Y}{\Omega'} \tag{267}$$

is a unit vector on the Bloch sphere, whose direction is determined by the ratio between the detuning $\delta\omega$ and the on-resonance frequency Ω. The Hamiltonian given by Eq. (266) thus describes the motion of a qubit in the Bloch sphere, which precesses around an effective axis along $\hat{\mathbf{n}}$ at a frequency Ω'.

In Fig. 9, we depict Rabi oscillations around the Bloch sphere about the $\hat{\mathbf{x}}$ axis, and plot oscillations for a superconducting qubit from 0–250 ns, in which we measure the excited state $|1\rangle$ population as a function of time. We find that the qubit coherently oscillates between the ground $|0\rangle$ and excited $|1\rangle$ states. The frequency of oscillation depends on the driving amplitude of the Rabi pulse, with higher amplitudes leading to faster oscillations. In general, one can drive a qubit at a frequency that is near-resonant and still observe oscillations, although off-resonant drives will not produce full contrast between $|0\rangle$ and $|1\rangle$ (see Sec. VI C 1). Therefore, the coarse measurements of frequency given by frequency-domain spectroscopy are generally sufficient to probe the coherent nature of a qubit via Rabi oscillations. Beyond being a fundamental test of qubit control, Rabi oscillations are also important for single-qubit gates, which are typically calibrated using resonant Rabi pulses. This requires the precise characterization of qubit frequencies, which is the topic of the following section.

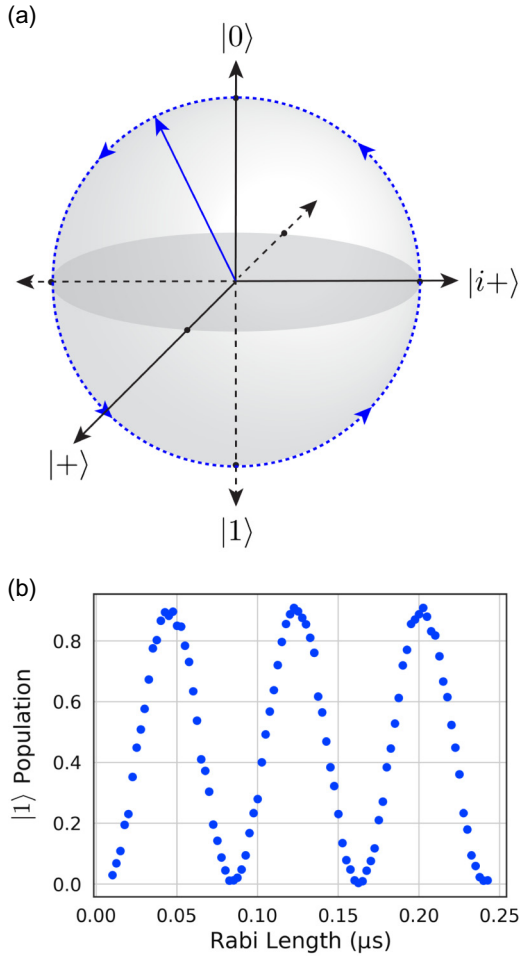


FIG. 9. Rabi oscillations. (a) Rabi oscillations are performed by driving a qubit (represented by a blue arrow on the Bloch sphere) coherently between the ground state $|0\rangle$ and excited state $|1\rangle$. (b) Rabi oscillations of a superconducting qubit. The qubit is driven on resonance and the probability of finding the qubit in the excited state $|1\rangle$ is measured as a function of time. The offset from 1.0 of the peaks of the oscillations is due to finite readout error of the excited state and/or an off-axis rotation, in which case the oscillations would not reach full contrast between $|0\rangle$ and $|1\rangle$.

C. Time-domain spectroscopy

Characterizing qubit frequencies is a fundamental component of performing high-fidelity qubit operations. Imperfect frequency calibrations or off-resonant qubit drives will result in coherent phase errors. These errors can be modeled with a small modification to an arbitrary single-qubit density matrix [Eq. (45)],

$$\rho = \begin{pmatrix} |\alpha|^2 & \alpha\beta^* e^{i\delta\omega t} \\ \alpha^*\beta e^{-i\delta\omega t} & |\beta|^2 \end{pmatrix}, \tag{268}$$

where we have added an explicit phase term $\exp(\pm i\delta\omega t)$, and $\delta\omega = \omega_q - \omega_d$ is the detuning between the qubit frequency and the drive frequency, which determines the rotating frame. This phase term accounts for a drive frequency that is off-resonant from the qubit frequency, in which case the qubit will precess in the rotating frame. In most cases, single-qubit quantum gates are designed to drive qubits on resonance; therefore, characterizing and correcting any off-resonant phase errors is important for gate calibration. While coarse measurements of qubit frequencies can be performed using classical frequency-domain spectroscopy, introduced in Sec. VI A, fine-tuned measurements of qubit frequencies require quantum-based protocols. In this section, we outline two fundamental characterization methods for measuring the detuning in qubit drive frequencies. The first is based on Rabi oscillations, outlined in the previous section, and the second introduces an important interferometric method known as *Ramsey spectroscopy*. Together, these methods are often referred to as *time-domain spectroscopy*, because they are generally implemented by driving a qubit at a given frequency for specific duration of time.

1. Rabi chevron

A basic method for finding the resonant frequency of a qubit is to Rabi drive the qubit across a range of different frequencies and measure the resulting period of the Rabi oscillations. As shown in the Rabi oscillation formalism, the effective Rabi frequency Ω' increases with larger detuning $\delta\omega$, resulting in shorter oscillation periods. Moreover, the amplitude of oscillations also decreases with larger detuning. Therefore, when the Rabi drive is on resonance with the qubit, both the amplitude and period of oscillations will be at peak value. When sweeping over a large range of detunings around the expected qubit resonance and measuring the resulting Rabi oscillations, a chevronlike pattern is produced, as shown in Fig. 10. Here, we observe that the Rabi oscillations are largest closest to the middle of the frequency sweep, and that the period and amplitude of oscillations slowly dies off at larger detunings, suggesting that the qubit drive is already near resonant with the qubit frequency. By finding the detuning at which peak oscillations occur, one is able to accurately characterize the drive frequency needed to perform resonant operations on the qubit.

2. Ramsey spectroscopy

Ramsey spectroscopy [220], or Ramsey interferometry, is a precise method for characterizing qubit frequencies. In a typical Ramsey experiment, a qubit is prepared in a superposition state (via a $X_{\pi/2}$ or $Y_{\pi/2}$ pulse) and allowed to evolve naturally for some amount of time t, after which the qubit is mapped back to the computational basis via the same gate used to prepare the state, and subsequently measured [see Fig. 11(a)]. Within the rotating frame of the qubit drive, defined by the frequency $f_d = \omega_d/2\pi$, any finite detuning between the qubit frequency and the drive frequency $\delta\omega$ will result in a qubit state that precesses along the equator of the Bloch sphere. Measuring the qubit in the computational basis for different times will result in a sinusoidal oscillation between $|0\rangle$ and $|1\rangle$, much like measurements of Rabi oscillations. However, in this case the oscillations are not caused by coherent driving between $|0\rangle$ and $|1\rangle$, but rather by coherent precession between $|+i\rangle$ and $|-i\rangle$ due to a frequency detuning, which is subsequently mapped back to the computation basis.

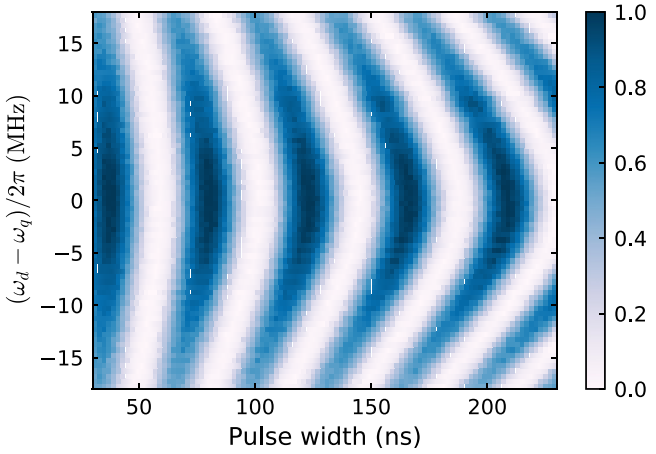


FIG. 10. Rabi chevron. A superconducting qubit is driven across a range of different frequency detunings, producing a chevron-shaped interferometry pattern. Each horizontal line in the above plot is an individual Rabi experiment, with the blue regions depicting the peaks of the oscillations and the white regions depicting the troughs. The qubit frequency can be found by extracting the drive frequency at which the period of oscillations is the largest; this corresponds to the apex of the chevron pattern.

In addition to the sinusoidal oscillations caused by any frequency detuning, a qubit in a superposition state will also experience stochastic noise along the longitudinal axis of the qubit due to interactions with the environment, causing the qubit frequency to fluctuate in time, which results in a Bloch vector that precesses both forwards and backwards in the rotating frame. This process is known as *pure dephasing* [see Fig. 4(b) and Sec. III B], which results in the depolarization of the Bloch vector towards the polar axis of the Bloch sphere. Pure dephasing will lead to the exponential decay of the Ramsey oscillations as a function of time. Thus, measurements of Ramsey spectroscopy are typically fit to an exponential cosine function, for which the probability of measuring the qubit in the excited state is given by

$$P_{|1\rangle}(t) = e^{-\Gamma_2 t}\cos(\omega_m t)\,, \tag{269}$$

where Γ_2 is the rate at which the qubit loses phase coherence, and ω_m is the measured frequency of oscillations.

The dephasing rate Γ_2 places a limit on the duration of time over which Ramsey oscillations can be observed. Therefore, if Γ_2 is large and/or $\delta\omega$ is small, it may be difficult to fit the sinusoidal component of Eq. (269) to the observed data. For this reason, it is common to drive the qubit at an intentionally large detuning, such that $\delta\omega = (\omega_q - \omega_d) + \Delta$, where Δ is an additional *artificial* detuning which has been added to the natural detuning. This enables one to fit the sinusoidal component to the observed Ramsey oscillations at short timescales even if the natural detuning is small.

In Fig. 11(b), we plot the Ramsey oscillations of a superconducting qubit for four different artificial detunings, $\Delta = \{-2, -1, 1, 2\}$ MHz. We observe that $P_{|1\rangle}$ oscillates sinusoidally, with only a small exponential component visible due to the short time span of the measurements (2 μs). By extracting the measured frequency of oscillation ω_m, we can compute the *measured detuning* (i.e., the difference between the measured frequency and the artificial detuning, $\omega_m - \Delta$) for each artificial detuning. In Fig. 11(c), we plot the measured detuning versus the artificial detuning.

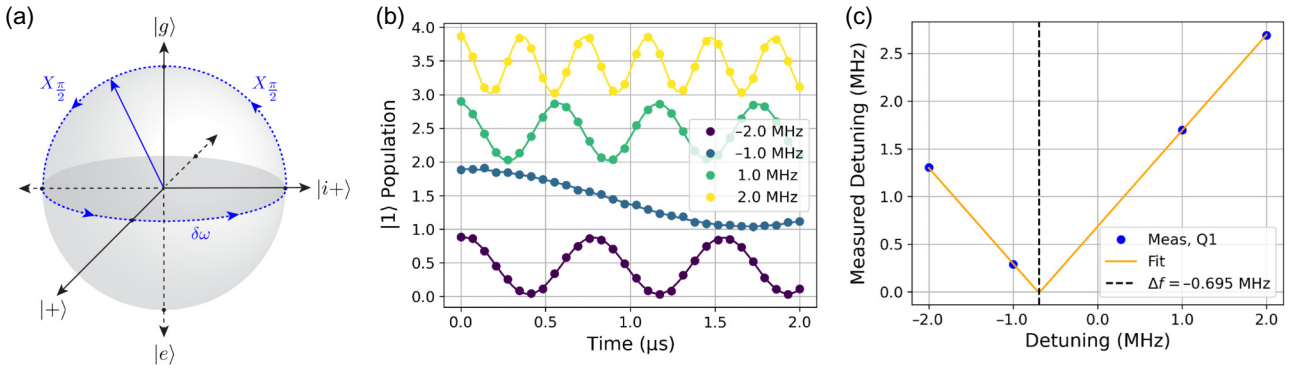


FIG. 11. Ramsey spectroscopy. (a) Ramsey oscillations around the Bloch sphere. A qubit prepared in a superposition state with an $X_{\pi/2}$ pulse will precess along the equator at a frequency given by $\delta\omega$, where $\delta\omega = \omega_q - \omega_d$ is the detuning between the qubit frequency ω_q and the frequency of the rotating frame ω_d. After some time t, another $X_{\pi/2}$ gate is performed and the qubit is measured in the computational basis. (b) Ramsey spectroscopy of a superconducting qubit (Q1) is performed for four different artificial detunings ($\Delta = \{-2, -1, 1, 2\}$ MHz). The resulting Ramsey oscillations are plotted and spaced vertically apart for visual clarity. The data are fit to an exponential cosine function, from which the frequency of oscillations can be extracted. (c) Measured detunings are extracted from the frequency fits from (b) and plotted as a function of the artificial detunings. The data is then fit to an absolute value function, with the vertex of the fit representing the actual frequency detuning of the qubit drive. From these data, it was found that the qubit was detuned 695 kHz from the drive frequency.

By choosing artificial detunings above and below where we expect the true qubit frequency to reside, we can fit the measured detuning to an absolute value curve, with the vertex of the fit representing the detuning between the qubit frequency and the drive frequency. For the data in Fig. 11(c), we find that the qubit frequency was detuned 695 kHz below the qubit drive.

D. Qubit coherence

There are two important characteristic timescales that define the *coherence* of a qubit. The first timescale—called T_1—describes how long a qubit will remain in an excited state before it decays to the ground state. The second characteristic timescale—called T_2—describes how long a qubit can maintain phase coherence in a superposition state. The characterization of these timescales are important, as they place fundamental limits on the gate fidelities achievable for each qubit, as well as fundamental limits on the time within which one can perform useful computations on a quantum processor.

In general, different types of qubits can have drastically different coherence times. For example, while coherence times of superconducting qubits ranging from ∼100 μs to 1 ms are considered quite long [221], atomic-based systems such as neutral atoms or trapped ions can exhibit drastically longer coherence times, ranging from seconds to even hours [222]. However, gate times can also differ by orders of magnitude between different platforms, typically ranging from tens of nanoseconds on superconducting systems to tens of milliseconds on atomic systems. Therefore, in the context of gate-based quantum computing, one should consider the relative number of gates that can be implemented within the coherence times of qubits on a quantum processor, as this determines the maximum circuit depth achievable for the processor.

1. Energy relaxation: T_1

A qubit in an excited state will eventually decay to the ground state due to energy relaxation, such as spontaneous emission (see Sec. III C). The characteristic timescale for thermalization—termed T_1—is defined by the longitudinal relaxation rate Γ_1 [Eq. (152)],

$$T_1 \equiv \frac{1}{\Gamma_1}. \tag{270}$$

T_1 is the $1/e$ decay constant for energy relaxation, whereby with some probability $p(t) = 1 - \exp(-t/T_1)$ at time t a qubit in the excited state will thermalize to the ground state. For an arbitrary single-qubit state, this process can be modeled in the density matrix formalism:

$$\rho = \begin{pmatrix} 1 + (|\alpha|^2 - 1)e^{-t/T_1} & \alpha\beta^* e^{i\delta\omega t} \\ \alpha^*\beta e^{-i\delta\omega t} & |\beta|^2 e^{-t/T_1} \end{pmatrix}, \tag{271}$$

where we note that as $t \longrightarrow \infty$, $\rho_{00} = 1 + (|\alpha|^2 - 1)e^{-t/T_1} \longrightarrow 1$ and $\rho_{11} = |\beta|^2 e^{-t/T_1} \longrightarrow 0$.

To measure the T_1 time of a qubit, the qubit is prepared in the $|1\rangle$ state using an X_π pulse and then measured after waiting some time t. By repeating this process for many different times, the measured data is fit to an exponential decay function $A\exp(-t/T_1)$, from which T_1 can be extracted. In Fig. 12, we plot the T_1 characterization curve for a superconducting transmon qubit, and find that it has a T_1 time of 102.0 (1.5) μs.

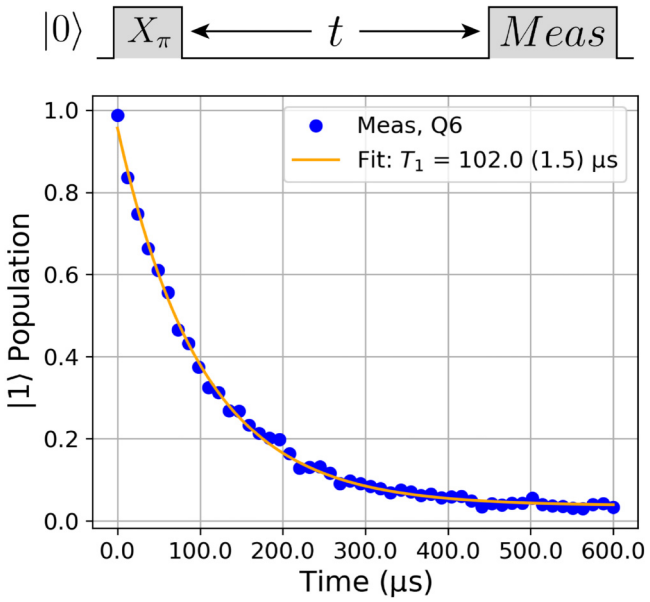


FIG. 12. Qubit T_1 characterization. Exponential decay curve for a superconducting qubit (labeled Q6) prepared in the excited state and measured after a variable amount of time. The raw data (blue points) corresponds to the ensemble probabilities of the qubit being measured in the $|1\rangle$ state after each waiting period, and orange curve is the exponential fit to the data. From this fit, we can extract a characteristic time of $T_1 = 102.0(1.5)$ μs for this qubit.

2. Phase decoherence: T_2

A qubit prepared in an superposition state will eventually experience phase decoherence due to both energy relaxation and pure dephasing (see Sec. III B). The characteristic timescale for phase decoherence—termed T_2—is defined by the transverse relaxation rate Γ_2 (Eq. (269)),

$$T_2 \equiv \frac{1}{\Gamma_2} = \left(\frac{\Gamma_1}{2} + \Gamma_\phi\right)^{-1}, \tag{272}$$

where Γ_ϕ is the rate of pure dephasing. Here, we see that in the limit of no pure dephasing (i.e., $\Gamma_\phi = 0$), the timescale for phase coherence is determined by the timescale for energy relaxation, with $T_2 = 2T_1$. This reflects the fact that T_1 events erase all phase knowledge of the qubit state, limiting the maximum length of time that a qubit can maintain phase coherence.

T_2 is the $1/e$ decay constant for phase decoherence, whereby with some probability $p(t) = 1 - \exp(-t/T_2)$ at time t a qubit in a superposition state will depolarize toward the polar axis. Equation (271) can be modified to include phase decoherence,

$$\rho = \begin{pmatrix} 1 + (|\alpha|^2 - 1)e^{-t/T_1} & \alpha\beta^* e^{i\delta\omega t} e^{-t/T_2} \\ \alpha^*\beta e^{-i\delta\omega t} e^{-t/T_2} & |\beta|^2 e^{-t/T_1} \end{pmatrix}, \tag{273}$$

where we have added e^{-t/T_2} to the off-diagonal terms to account for phase decoherence as $t \longrightarrow \infty$. In the long-time limit, all terms converge to zero except ρ_{00}. Equation (273) is known as the Bloch-Redfield model of two-level systems [223].

The phase decoherence time T_2 can be measured using Ramsey spectroscopy (Sec. VI C 2). First, the qubit is prepared in a superposition state, then allowed to naturally dephase along the equator for a variable amount of time, after which the resulting state is rotated back to the computational basis and subsequently measured [see Fig. 11(a)]. In a Ramsey experiment, one should observe decaying oscillations between $|0\rangle$ and $|1\rangle$. By fitting the data to a decaying sinusoid [Eq. (269)], T_2 can be determined directly from the exponential fit parameter, $T_2 = 1/\Gamma_2$. If both the drive-qubit detuning and the dephasing rate are small, then it can be difficult to fit the observed data to Eq. (269). In this case, it is convenient to add an artificial detuning to the drive (see Fig. 11) such that the data can be accurately fit to a decaying sinusoid. The dephasing time measured using Ramsey spectroscopy is typically written as T_2^* to denote that it is sensitive to inhomogeneous low-frequency noise (see the discussion in Sec. III B). In Fig. 13(a), we plot the T_2^* characterization curve for a superconducting transmon qubit, and find that it has a T_2^* time of 140.0 (5.3) μs. We observe that it is less than $2T_1$ of the same qubit (see Fig. 12), indicating the presence of pure dephasing.

Ramsey measurements are generally sensitive to low-frequency noise. Here, low-frequency noise is defined to be quasistatic over the timescale of an experiment, but can vary from experiment to experiment. Therefore, it is possible to "echo" away the effect of the noise using a Hahn-echo pulse [86]. Hahn-echo experiments are identical to Ramsey experiments in the state preparation and measurement, but halfway through the experiment an X_π pulse is applied to the qubit. This reverses the effects of inhomogeneous broadening caused by the quasistatic noise, effectively refocusing the Bloch vector. By performing a Hahn-echo experiment for different durations of time, one can fit the measurements to an exponential function whose exponential fit parameter determines the T_2 time of the qubit with an echo pulse—denoted T_{2E}. In Fig. 13(b), we plot the T_{2E} characterization curve for a superconducting transmon qubit, and find that it has a T_{2E} time of 160.0 (6.1) μs. While this is longer than the T_2^* time of the qubit, it does not saturate the $T_2 = 2T_1$ limit. This indicates the presence of not only low-frequency quasistatic noise (since $T_{2E} > T_2^*$), but also high-frequency noise, which likely varies over the timescale of the experiment.

E. Phase estimation

While *continuous* Rabi, Ramsey, and Hahn-echo experiments are useful for learning basic properties of qubits,

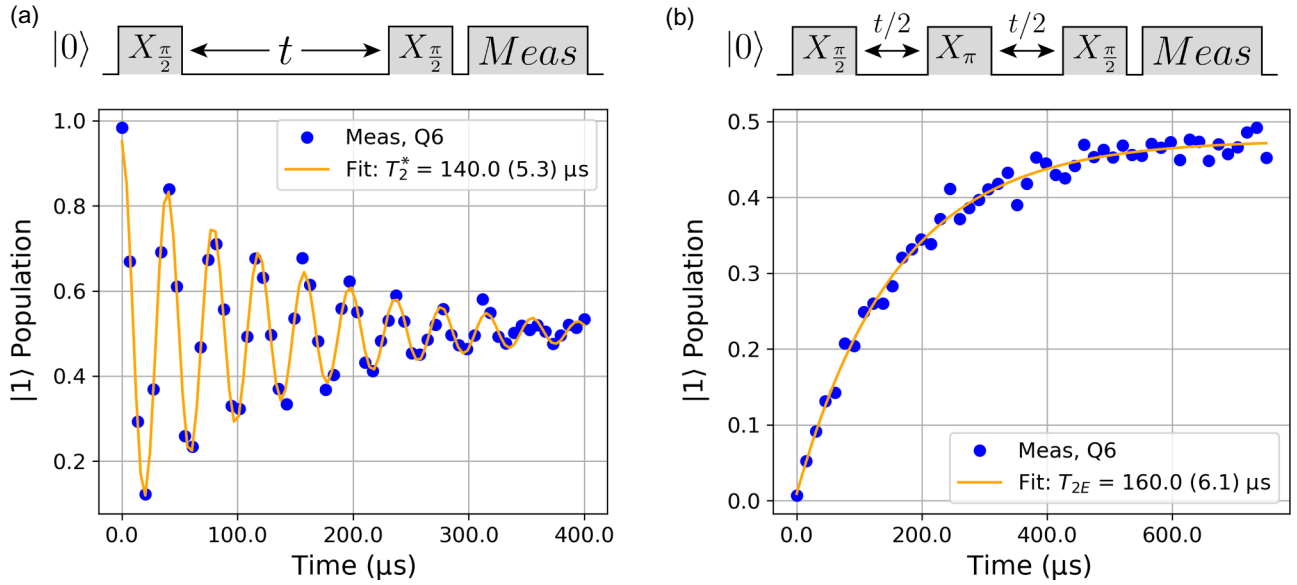


FIG. 13. Qubit T_2 characterization. (a) Decaying sinusoidal curve for a superconducting qubit (labeled Q6), which has been prepared in a superposition state using an $X_{\pi/2}$ pulse, allowed to precess along the equator for some time t, after which it is mapped back to the computational basis with a final $X_{\pi/2}$ pulse and subsequently measured. From the fit of the data, we extract a characteristic time of $T_2^* = 140.0(5.3)$ µs for this qubit. (b) Hahn-echo experiment for Q6. The sequence is identical to the Ramsey sequence in (a), but an X_π pulse is added to the middle of the experiment to echo away the effects of low-frequency noise. From the exponential fit, we extract a characteristic time of $T_{2E} = 160.0(6.1)$ µs for this qubit.

they do not provide detailed information about the performance of quantum gates in a gate-based setting. However, by performing *discrete* Rabi and Ramsey experiments composed of defined quantum logic operations (e.g., $X_{\pi/2}$ or $Y_{\pi/2}$ gates), one can learn detailed information about the underlying gates themselves. This is the goal of a set of characterization methods under the broad term *phase estimation* [224].

As an ideal n-qubit quantum gate implements a unitary in $\mathsf{SU}(2^n)$, we may think of any such gate as implementing some manner of *rotation* of vectors in $d = 2^n$ dimensional Hilbert space. Implementing incorrect rotation angles is a primary source of coherent error in quantum hardware (see Sec. III A); accurate characterization of such angles is necessary for the calibration of high-quality gates. The task of specifically estimating a gate's rotation angle is called phase estimation, and it is the task we concern ourselves with in this subsection.

While there exist multiple protocols for phase estimation, here we review a particular flavor of it called *robust phase estimation (RPE)* [225]. RPE may be thought of in some sense as an interpolation between the aforementioned Rabi and Ramsey oscillations and the gate set tomography (GST) protocol, discussed in Sec. VII D. Like Rabi oscillations, RPE estimates one particular Hamiltonian parameter of a gate operation, but like GST, it uses a set of circuits with logarithmically spaced depths, allowing it to learn that parameter with Heisenberg-like accuracy.

Without loss of generality, we may consider the task of estimating the phase θ from a single-qubit gate $U = e^{-i\theta\sigma_j/2}$, with $\theta \in [-\pi, \pi]$ and where σ_j may be taken to be any Pauli matrix [226]. If we simply apply the gate U to a uniform superposition of its two eigenstates and then perform a projective measurement onto that same superposition, the probability P_c that the system is projected onto that same superposition is given by

$$P_c = |\langle +|U|+\rangle|^2 = \frac{1+\cos\theta}{2}, \tag{274}$$

where $|+\rangle$ denotes the uniform superposition of the two eigenstates of U [by analogy with the standard definition of $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$]. Similarly, if we instead perform a projective measurement onto the $|i+\rangle$ state (where we have put a relative phase of i between the two eigenstates), the probability of measuring in the $|i+\rangle$ state is

$$P_s = |\langle i+|U|+\rangle|^2 = \frac{1+\sin\theta}{2}. \tag{275}$$

By repeating both of these experiments many times to build up approximations of P_c and P_s—which we will denote as $\hat{P}_c$ and $\hat{P}_s$, respectively—we can estimate θ by $\hat{\theta}$:

$$\hat{\theta} = \arctan 2(2\hat{P}_s - 1, 2\hat{P}_c - 1), \tag{276}$$

where $\arctan 2$ is the arc-tangent function which accounts for branch cuts by tracking the signs of its two arguments.

While one can, in principle, estimate θ in this manner, this approach suffers from two problems. First, unwanted error terms (e.g., decoherence effects, SPAM errors, etc.) can pollute $\hat{P}_c$ and $\hat{P}_s$, corrupting $\hat{\theta}$. Second, even if such errors are *not* present, the accuracy of the estimate $\hat{\theta}$ is at the standard quantum limit, i.e., if N repetitions (shots) of each of the two circuits are taken, then the uncertainty in $\hat{\theta}$ scales as $1/\sqrt{N}$. For learning θ to high precision, this approach becomes very expensive.

RPE solves both of these problems, as we describe in the rest of this subsection. Instead of only using circuits with just a single repetition of U, RPE replaces the single instances of U in Eqs. (274) and (275) with k repetitions of U, for $k \in \{1, 2, 4, \ldots, k_{\max}\}$, giving an RPE experiment a total of $2(1 + \log_2 k_{\max})$ circuits. Thus, the target probability distributions that an RPE experiment attempts to sample from look unsurprisingly similar to Eqs. (274) and (275). However, as we noted above, Eqs. (274) and (275) do not take into account other errors, which could perturb the distributions we wish to sample from (and estimate). If we denote such additive perturbations by $\delta_{k,c}$ and $\delta_{k,s}$ for the $P_{k,c}$ and $P_{k,s}$ distributions, respectively, then we find that the *actual* probability distributions an RPE experiment samples from are given by

$$P_{k,c} = \frac{1 + \cos(k\theta)}{2} + \delta_{k,c}\,, \tag{277}$$

$$P_{k,s} = \frac{1 + \sin(k\theta)}{2} + \delta_{k,s}\,. \tag{278}$$

As detailed in Refs. [225,227], built into RPE is a robustness against such additive errors. As long as $\max_{\{i\in\{c,s\},k\}} |\delta_{k,i}| < \sqrt{\frac{3}{32}} \approx 30.6\%$, RPE can still successfully estimate θ. In fact, if that constraint on the additive errors is satisfied for all k up to some $k_{\max}$, then the root mean square (RMS) error of RPE's estimate of θ will be no greater than $\pi/(2 \cdot k_{\max})$. Thus, RPE yields an estimate of θ that is Heisenberg-limited in its accuracy (up to decoherence), allowing θ to be estimated extraordinarily efficiently. For example, it was shown in Ref. [228] that RPE could be used to learn a single-qubit gate's phase to within 4×10^{-4} radians, with only 176 total experimental samples.

We now turn to discussing *how* RPE constructs an estimate of θ from its experimental estimates $\hat{P}_{k,c}$ and $\hat{P}_{k,s}$. For a given "generation" of circuits corresponding to k repetitions of U, we can estimate $k\theta$ according to

$$k\hat{\theta} = \arctan 2(2\hat{P}_{c,k} - 1, 2\hat{P}_{k,s} - 1) \mod 2\pi\,. \tag{279}$$

Because sinusoids exhibit periodicity, we cannot learn θ from just Eq. (279). In other words, there are multiple values of $\hat{\theta}$ that satisfy Eq. (279). However, we can use successive generations of RPE data to learn θ iteratively. For each successive generation, the angular space in which $\hat{\theta}$ can fall is cut in half (thus allowing the uncertainty to shrink by a factor of two with every generation, yielding Heisenberg-like scaling). To begin, we start with $k = 1$, and compute an initial estimate of $\hat{\theta}_1$, given by Eq. (276). For each subsequent generation, we compute

$$\tilde{\theta}_k = \arctan 2(2\hat{P}_{c,k} - 1, 2\hat{P}_{k,s} - 1)/k\,. \tag{280}$$

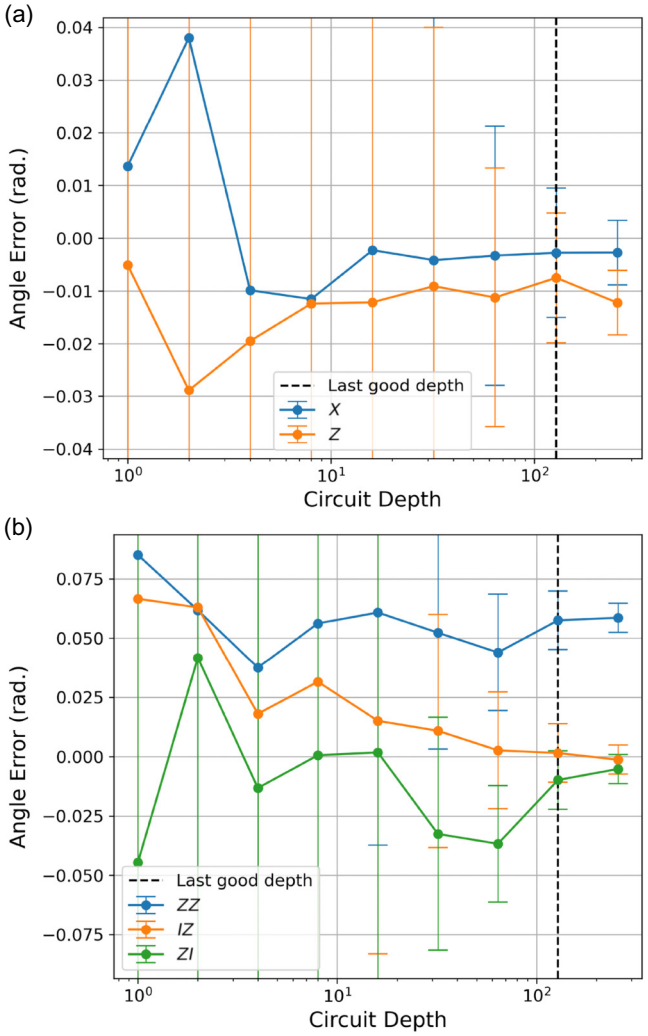


FIG. 14. Robust phase estimation. (a) RPE of an $R_x(\pi/2)$ gate. The X error (i.e., over or under rotation) and Z error (i.e., axis angle error) in the gate are plotted above. Both the X and Z errors are equal to zero (up to uncertainty). (b) RPE of a CZ gate. Rotation errors on the IZ, ZI, and ZZ Hamiltonian generators of the gate are plotted above. RPE is able to verify that the errors on the IZ and ZI terms are zero (up to uncertainty), but there remains an error on the ZZ entangling term of approximately $0.06 \pm \frac{\pi}{512} \approx 0.06 \pm 0.006$ radians. The error bars in the plot represent the $\pi/(2 \cdot k_{\max})$ upper bound on the estimates' RMS error (we omit displaying the full error bar at lower circuit depths for visual clarity), and the most trusted estimate of the errors (i.e., "last good depth") from the RPE consistency checks are indicated by dashed vertical lines.

This quantity, $\tilde{\theta}_k$, is effectively the update to our previous estimate $\hat{\theta}_{k-1}$. However, in order for the new estimate to be in the correct branch, we must add (or subtract) $2\pi/k$ to (or from) $\hat{\theta}_k$ until the resulting quantity falls within $\hat{\theta}_{k-1} \pm \pi/k$. This resulting quantity is then taken to be $\hat{\theta}_k$.

In Fig. 14(a), we plot the results of an RPE experiment performed on a single-qubit $R_x(\pi/2)$ gate implemented on a superconducting qubit and analyzed using the python package `pyRPE` [229]. As discussed in Refs. [225,228], small modifications can be made to the standard RPE routine to learn both the error in rotation angle and the axis misalignment of a gate relative to the target operation [i.e., an ideal $R_x(\pi/2)$ gate]. We plot both of these errors for the $R_x(\pi/2)$ gate in Fig. 14(a), where the X error refers to an over or under rotation, and the Z error refers to a phase (i.e., axis) error on the gate. The error bars in the plot represent the $\pi/(2 \cdot k_{\max})$ upper bound on the phase estimates' RMS error, and the most trusted estimate of the errors are indicated from the RPE consistency checks. Both the X and Z errors on the $R_x(\pi/2)$ are zero, up to uncertainty.

Additionally, RPE may be used to characterize two-qubit rotation angles, with only minor modifications required. Ref. [230] demonstrated RPE on a CZ gate that acts on pair of superconducting qubits. As, up to a global phase,

$$\mathrm{CZ} = e^{\frac{-i}{2}\left(\frac{\pi}{2}ZI + \frac{\pi}{2}IZ - \frac{\pi}{2}ZZ\right)}, \tag{281}$$

RPE can be used to learn each of the coefficients of the ZI, IZ, and ZZ terms in the Hamiltonian which generates the CZ gate. The results, summarized in Fig. 14(b), show that while the errors on the IZ and ZI phases are zero (up to uncertainty), the ZZ entangling phase has an angle error of approximately 0.06 radians. Because RPE is rather inexpensive to perform, the aforementioned experiment cost fewer than 60 distinct circuits, and its error estimates can be used to (re)calibrate the gates being characterized using parameter sweeps or closed-loop optimization [231].

VII. TOMOGRAPHIC RECONSTRUCTION

Tomographic reconstruction methods (a.k.a. "tomography") estimate all aspects of an object by probing it along several axes and combining the results. In QCVV, tomography is used to estimate the mathematical object representing a quantum logic operation—a quantum state, process, or measurement. Tomography protocols estimate the *entire* density matrix, transfer matrix, or POVM. This requires more experimental data and analytic effort than just estimating a few properties of the object, but yields more diagnostic power. A complete tomographic characterization reveals *all* interesting properties of a logic operation, enabling debugging of quantum hardware and predicting the behavior of operations *in situ*.

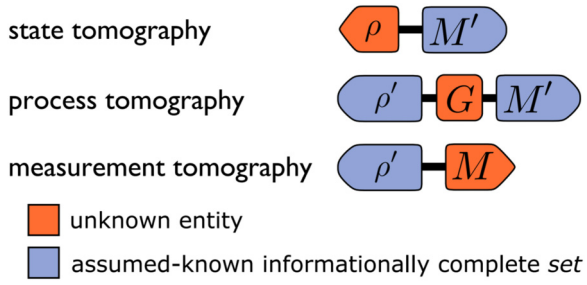


FIG. 15. Tomographic reconstruction. The structure of the circuits required for state, process, and measurement tomography are shown. Each of these protocols reconstructs an initially unknown quantum operation (a state ρ, process G, or POVM M) by combining that operation into simple circuits with a set of *known* complementary operations. The known operations form a reference frame for estimation of the unknown operation's matrix elements. If the reference frame is *informationally complete*, then all matrix elements of the unknown operation can be learned. We denote reference frame operations with "primed" symbols (ρ' and M') to indicate that they are *effective* (rather than *native*) state preparations and measurements, usually implemented by applying gate operations after or before a native state preparation or measurement. These techniques are limited by systematic errors stemming from imperfect *a priori* knowledge of the ρ' and M'. Figure and caption reproduced with permission from Ref. [156].

Quantum tomography methods are among the oldest characterization tools. Quantum state tomography (Sec. VII A) appears in the literature as early as 1968 [232], and quantum process tomography (Sec. VII B) dates to 1997 [233]. Traditional methods such as quantum state and process tomography are still widely used to this day, but they are unreliable in the presence of imperfect state preparation and measurement (SPAM) [194]. Self-consistent tomographic methods like gate set tomography (Sec. VII D) avoid this problem, and have superseded state and process tomography in contexts that require reliability.

In this section, we provide overviews of the following tomographic reconstruction methods:

(a) *Quantum state tomography* (Sec. VII A). Quantum state tomography is designed to reconstruct an unknown quantum state ρ (i.e., density matrix) by performing an informationally complete set of measurements $\{M'_j\}$ on many identical copies of ρ (see Fig. 15).

(b) *Quantum process tomography* (Sec. VII B). Quantum process tomography is designed to reconstruct an unknown quantum operation (e.g., a gate G) by applying many identical implementations of G to an informationally complete set of distinct states $\{\rho'_i\}$ and performing an informationally complete set of measurements $\{M'_j\}$ on many identical copies of each $G[\rho'_i]$ (see Fig. 15).

(c) *Quantum measurement tomography* (Sec. VII C). Quantum measurement tomography is designed to reconstruct an unknown POVM M by applying it to an informationally complete set of states $\{\rho_i'\}$ (see Fig. 15).

(d) *Gate set tomography* (Sec. VII D). Gate set tomography is designed to self-consistently reconstruct an entire *set* of quantum operations including at least one initialization (state), at least one measurement (POVM), and at least two logic gates (quantum processes)—i.e., a *gate set*—by running a wide range of circuits composed from those operations (see Fig. 20).

A. Quantum state tomography

The goal of *quantum state tomography (QST)* [232] is to accurately estimate the density matrix ρ representing a quantum state. This cannot be done using just one copy of the unknown state ρ, because no single measurement will reveal ρ, and measuring a system disrupts its unmeasured properties. State tomography therefore requires *many* (N) identically prepared systems. It is usually assumed, for simplicity, that these systems are identically and independently prepared, so that their joint state is $\rho^{\otimes N}$ for some unknown ρ. In real experiments, this assumption is only an approximation, and state tomography estimates the *average reduced density matrix* of the N sample systems.

State tomography is performed by performing an *informationally complete* measurement, or set of measurements, on the samples of ρ (see Fig. 15). Informational completeness is a property of a set of POVM effects $\{E_i\}$, and it does not matter whether those effects all come from a single POVM (i.e., $\{E_i\}$ is itself an experimentally performable POVM M') or a set of distinct measurements (e.g., $\{E_i\}$ is the union of the effects of several PVMs $\{M_1', M_2', \ldots\}$). A set of effects is informationally complete if (and only if) they span the vector space $\mathcal{B}(\mathcal{H})$ of operators. Since $\mathcal{B}(\mathcal{H})$ for a d-dimensional Hilbert space $\mathcal{H}$ is d^2-dimensional, a measurement or set of measurements must contain d^2 linearly independent effects to be informationally complete and enable state tomography. This is usually achieved by choosing at least $d+1$ PVMs (orthogonal bases), but can in principle be achieved with a single d^2-element POVM. Informationally complete sets are not all created equal—the accuracy of tomographic reconstruction is controlled by the condition number of $\{E_i\}$'s Gram matrix, and optimal accuracy is achieved by a *2-design* [185], such as a full set of mutually unbiased bases [234,235] or a symmetric informationally complete POVM [185,236].

The basic principle of state tomography is very simple. By repeating measurements many times, we estimate the probability $p(E_i)$ of each effect. The simplest estimator $\hat{p}(E_i)$ is the number of times E_i was observed divided by the number of times it could have occurred:

$$\hat{p}(E_i) = \frac{n_i}{N_i}. \tag{282}$$

By Born's rule,

$$p(E_i) = \mathrm{Tr}(E_i\rho) = \langle\langle E_i \,|\, \rho \rangle\rangle. \tag{283}$$

If the set $\{E_i\}$ spans $\mathcal{B}(\mathcal{H})$, then Eq. (283) defines a set of linear equations that can be solved uniquely for ρ [see Eq. (295) below]. This is state tomography.

To gain insight into how this is done in practice, consider the simplest case of a single-qubit state described by a 2×2 density matrix (see Sec. II B 1). A single projective measurement is not sufficient to determine ρ, even if we repeat it $N \longrightarrow \infty$ times. Measuring, for example, Z reveals only $p(|0\rangle\langle 0|)$ and $p(|1\rangle\langle 1|)$, from which we can deduce $\langle Z\rangle$ and $\langle \mathbb{I}\rangle$ but not $\langle X\rangle$ or $\langle Y\rangle$. To estimate ρ, we also need to learn $\langle X\rangle$ and $\langle Y\rangle$, because a single-qubit ρ is defined by a Bloch vector $\mathbf{r} \in \mathbb{R}^3$ in the Bloch ball [see Eq. (53) and Fig. 1].

We can obtain informationally complete data by dividing N samples of ρ into three groups, then measuring X on each sample in the first group, Y on the second group, and Z on the third group. From the resulting count data, $\{\langle X\rangle, \langle Y\rangle, \langle Z\rangle\}$ can all be estimated. Then the density matrix can be reconstructed as

$$\rho = \frac{1}{2}\left(I + \langle X\rangle X + \langle Y\rangle Y + \langle Z\rangle Z\right). \tag{284}$$

Often, the only native measurement is a Z-basis measurement. In this case, *effective* measurements of X and Y are performed by (1) rotating the qubit using a $R_Y(-\pi/2)$ or $R_X(\pi/2)$ operation (respectively), then (2) performing the native Z-basis measurement.

This informationally complete experiment can be described (as above) as a union of three distinct PVMs. But it can equally well be described as a *single* POVM with six outcomes, $M' = \left\{\frac{1}{3}|\psi_i\rangle\langle\psi_i|\right\}$, where $|\psi_i\rangle$ ranges over the six single-qubit Pauli eigenstates (i.e., the ± 1 eigenstates of X, Y, and Z). This POVM can be implemented by drawing a uniformly random number k from $1\ldots 3$ and performing the kth Pauli PVM (repeating both steps for each shot). State tomography experiments are sometimes described this way [237,238], representing a set of PVMs or POVMs as a single POVM, because it is simpler.

A straightforward extension of this single-qubit tomography protocol can be used to perform state tomography on an n-qubit system. To do single-qubit state tomography, we measured in three independent bases. For n-qubit state tomography, we need 3^n distinct "measurement configurations" or PVMs. Each measurement configuration corresponds to simultaneously measuring one of the three Paulis (X, Y, Z) on each of the n qubits. So, for example, state

tomography on $n = 2$ qubits uses $9 = 3^2$ distinct configurations each corresponding to measuring one of $\{X, Y, Z\}$ on the first qubit and (independently) one of $\{X, Y, Z\}$ on the second. Each of the nine measurements has four possible outcomes, which are represented by rank-1 POVM effects. For example, the four effects for measuring Z on both qubits are $\{|00\rangle\langle 00|, |01\rangle\langle 01|, |10\rangle\langle 10|, |11\rangle\langle 11|\}$. After performing this measurement on N samples, it is straightforward to estimate each effect's probability as $\hat{p}_{ij} = n_{ij}/N$ (e.g., $\hat{p}_{00} = n_{00}/N$). Born's rule says that $p_{ij} = \mathrm{Tr}(|ij\rangle\langle ij|\,\rho) = \langle ij|\,\rho\,|ij\rangle$. So, we can transform these estimated probabilities into estimates of the expectation values of three Pauli observables just by writing out those Paulis as linear combinations of POVM effects,

$$ZZ = |00\rangle\langle 00| - |01\rangle\langle 01| - |10\rangle\langle 10| + |11\rangle\langle 11|\,, \quad (285)$$

$$IZ = |00\rangle\langle 00| - |01\rangle\langle 01| + |10\rangle\langle 10| - |11\rangle\langle 11|\,, \quad (286)$$

$$ZI = |00\rangle\langle 00| + |01\rangle\langle 01| - |10\rangle\langle 10| - |11\rangle\langle 11|\,, \quad (287)$$

and then multiplying both sides by ρ and taking the trace:

$$\langle ZZ\rangle \simeq \frac{n_{00}}{N} - \frac{n_{01}}{N} - \frac{n_{10}}{N} + \frac{n_{11}}{N}\,, \quad (288)$$

$$\langle IZ\rangle \simeq \frac{n_{00}}{N} - \frac{n_{01}}{N} + \frac{n_{10}}{N} - \frac{n_{11}}{N}\,, \quad (289)$$

$$\langle ZI\rangle \simeq \frac{n_{00}}{N} + \frac{n_{01}}{N} - \frac{n_{10}}{N} - \frac{n_{11}}{N}\,. \quad (290)$$

The coefficients in this linear transformation from effect probabilities to Pauli expectation values correspond to the 2-bit Walsh-Hadamard transform [Eq. (407)], which can be defined for any n.

The expectation value of *every* n-qubit Pauli operator in the 4^n-element Pauli group $\mathbb{P}_n = \{I, X, Y, Z\}^{\otimes n}$ can be straightforwardly estimated from at least one of the 3^n measurement configurations defined above. Then, because the Pauli operators form a complete orthogonal basis for $\mathcal{B}(\mathcal{H})$, ρ can be reconstructed as

$$\rho = \frac{1}{2^n}\sum_{P\in\mathbb{P}_n} \langle P\rangle\, P\,. \quad (291)$$

In the example of $n = 2$ qubits,

$$\begin{aligned}\rho = \frac{1}{4}(&II + \langle IX\rangle IX + \langle IY\rangle IY + \langle IZ\rangle IZ\\ &+ \langle XI\rangle XI + \langle XX\rangle XX + \langle XY\rangle XY + \langle XZ\rangle XZ\\ &+ \langle YI\rangle YI + \langle YX\rangle YX + \langle YY\rangle YY + \langle YZ\rangle YZ\\ &+ \langle ZI\rangle ZI + \langle ZX\rangle ZX + \langle ZY\rangle ZY + \langle ZZ\rangle ZZ)\,. \quad (292)\end{aligned}$$

Figure 16 shows an experimental realization of this state tomography protocol to estimate a two-qubit Bell state.

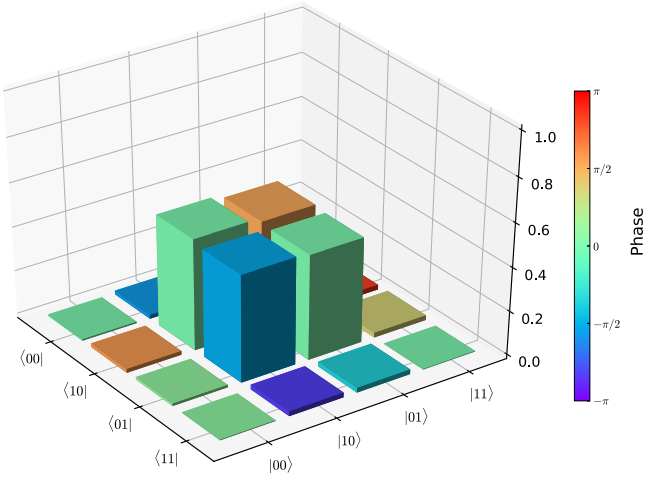


FIG. 16. Quantum state tomography. The estimated density matrix of a two-qubit Bell state that was experimentally realized by applying a $\sqrt{i\mathrm{SWAP}}$ gate to the initial state $|10\rangle$, and then reconstructed using quantum state tomography [218]. The tomographic estimate indicated that the desired Bell state was prepared with high fidelity ($F \approx 0.995$). Each bar represents a single element of the 4×4 density matrix, with its height indicating the matrix element's absolute value and its color indicating complex phase.

This simple n-qubit tomography protocol illustrates two valuable points. First, the n-qubit measurements described above are PVMs—their outcome effects are mutually orthogonal rank-1 projectors—but they are *not* measurements of individual Pauli observables. When we perform Z measurements simultaneously on both qubits, that is not "a measurement of ZZ." ZZ is a Pauli operator with only two eigenvalues ($+1$ and -1) that each label a two-dimensional eigenspace. A measurement of ZZ is described by a rank-2 POVM with two outcomes that yields exactly 1 bit of information. In contrast, measuring both qubits in Z yields four outcomes and 2 bits of information. This is a simultaneous measurement of multiple commuting Paulis, a.k.a. a *stabilizer* [239]. The simultaneously measured observables, ZI and IZ, generate a *stabilizer group* (a maximal Abelian subgroup of $\mathbb{P}_n$ containing 2^n commuting Pauli operators). Each of the 2^n Paulis in the stabilizer group can be written as a linear combination of the 2^n rank-1 projectors describing outcomes of the PVM measurement of the stabilizer. In this example, the Paulis $\{II, IZ, ZI, ZZ\}$ are linear combinations of $\{|00\rangle\langle 00|, |01\rangle\langle 01|, |10\rangle\langle 10|, |11\rangle\langle 11|\}$). A good estimate of the outcome probabilities of a stabilizer measurement is sufficient to estimate *all* the expectation values of the 2^n Paulis in the stabilizer. Note that some tomography theory papers do consider direct measurements of Pauli observables (which have two outcomes and reveal one expectation value) instead of stabilizers (which have 2^n outcomes and reveal $2^n - 1$ expectation values).

Second, the 3^n distinct stabilizer measurements are not mutually independent. Performing nine measurements that each have four outcomes should reveal $27 = 9(4-1)$ independent probabilities. But these 27 distinct probabilities are not linearly independent. For example, in one measurement configuration we measure the stabilizer $\{ZI, IZ\}$, while in another we measure $\{ZI, IX\}$. Each yields three Pauli expectation values, but they are not linearly independent because ZI appears in both stabilizers. In fact, ZI appears in the stabilizers of three different measurement configurations, whereas ZX only appears in one. As a result, this tomographic experiment yields more information about ZI than ZX, and therefore higher precision in the estimate of $\langle ZI \rangle$ than $\langle ZX \rangle$. In the n-qubit case, a Pauli that acts as $\mathbb{I}$ on k of the n qubits will appear in 3^k measurement configurations. As a result, this tomographic measurement protocol is both redundant (more measurement configurations than necessary) and heteroskedastic (some Pauli expectation values are estimated to much lower precision than others).

These features are unavoidable if each qubit is measured independently. There is a 1:1 correspondence between weight-n Paulis and measurement configurations, so removing even a single one of the 3^n measurement configurations breaks informational completeness (there is some Pauli $P \in \mathbb{P}_n$ whose expectation value cannot be estimated). However, by performing *entangling* n-qubit measurements—i.e., POVMs or PVMs whose effects are not tensor products of n single-qubit projectors—it is possible to construct a set of just $2^n + 1$ PVMs that measure mutually unbiased bases [234]. This set is informationally complete, minimal, and enables more accurate tomography than the local measurement described above.

State tomography can be performed on any quantum system, e.g., d-dimensional systems where $d \neq 2^n$. The basic principle is the same: (1) define a set of measurements whose effects $\{E_i\}$ span $\mathcal{B}(\mathcal{H})$; (2) perform those measurements on N samples of the unknown ρ; and (3) estimate ρ by inverting Born's rule. The main complications are technical. The Pauli operators cannot be used, and the available operator bases for d-dimensional qudits are less convenient (see Appendix B 4). Mutually unbiased bases are not known (or believed) to exist unless $d = D^n$, where D is prime. More details can be found in Appendix D 2 a.

1. Maximum likelihood estimation

The simple description of quantum state tomography in the previous section glosses over some key (if nonobvious) points:

(1) What should be done if we have measured *more than* d^2 distinct observables or POVM effects? In this case, the equations in Eq. (283) will *overconstrain* ρ and may have no solution.

(2) What should be done if solving Eq. (283) yields an estimated state $\hat{\rho}$ that is not positive semidefinite?

Both of these issues arise because finite-sample fluctuations (a.k.a. *shot noise*) [240] cause $\hat{\rho}$ to fluctuate randomly around the true ρ. We have ignored these fluctuations so far, implicitly assuming that the *estimated* value of any observable (e.g., $\langle Z \rangle$) is equal to its true value. But this is not true in practice. As a result, tomography is a *statistical* problem. The two issues highlighted above are solved by reformulating tomography not as a set of linear equations, but as a statistical inference problem.

The easiest way to address these issues is to treat Eq. (283),

$$p(E_i) = \mathrm{Tr}[E_i \rho] = \langle\langle E_i \,|\, \rho \rangle\rangle \,,$$

not as an exact linear inversion problem, but as a least-squares problem. These equations can be written in matrix form, by arranging the effect probabilities into a column vector $\vec{p} = [p(E_1), p(E_2), \ldots]^{\mathrm{T}}$ and stacking the vectorized effects into a matrix

$$T = \begin{pmatrix} \langle\langle E_1 | \\ \langle\langle E_2 | \\ \vdots \end{pmatrix} \tag{293}$$

so that

$$\vec{p} = T|\rho\rangle\rangle \,. \tag{294}$$

Now, if these equations have a unique solution $\hat{\rho}$, it is given by

$$|\hat{\rho}\rangle\rangle = T^{-1}\vec{p} \,. \tag{295}$$

This is linear inversion state tomography, in a single equation.

If we have measured more than d^2 observable probabilities, then T will not be square, and therefore not invertible. But if we admit that the estimated probabilities $\hat{p}(E_i)$ will fluctuate around the true probabilities, then we can reformulate Eq. (294) as a least squares problem and seek the $|\hat{\rho}\rangle\rangle$ that minimizes

$$\|\vec{p} - T|\rho\rangle\rangle\|_2^2 = \sum_i \left[\hat{p}(E_i) - \mathrm{Tr}(E_i\hat{\rho})\right]^2 . \tag{296}$$

This actually has a closed-form solution, in terms of the *Moore-Penrose pseudoinverse* of the matrix T:

$$|\hat{\rho}\rangle\rangle = T^{+}\vec{p} \,, \tag{297}$$

where the pseudoinverse is defined as

$$T^{+} \equiv (T^{\mathrm{T}}T)^{-1}T^{\mathrm{T}}\,. \tag{298}$$

The second issue (what if $\hat{\rho} \not\geq 0$) can also be solved by reformulating inversion as an ordinary least-squares problem, by *constraining* the optimization of Eq. (296) to positive semidefinite $\rho \geq 0$. This is a tractable convex optimization problem. It can be solved by *projecting* the unconstrained linear inversion estimate onto the nearest positive semidefinite state [241], an algorithm known as *projected least squares (PLS)* [242,243].

However, these *least-squares tomography* estimators are ad hoc solutions, and not optimal in any sense except simplicity. They can be seen as approximations to a statistically well-motivated approach called *maximum likelihood estimation (MLE)* [244,245]. MLE is a simple, widely used method for statistical inference—i.e., the estimation of unknown parameters from data—in which the estimated values of the unknown parameters are the ones that maximize the probability of observing the data that were actually observed. The probability of the observed data, as a function of the unknown parameters, is called the *likelihood function*:

$$\mathcal{L}(\theta) = \Pr(D_{\mathrm{observed}}|\theta)\,. \tag{299}$$

Here, θ is a vector of parameters whose values we would like to estimate, and D_{observed} is the actual data that have been observed. The likelihood function is a compressed, efficient representation of the information that D_{observed} provides about the unknown θ. The maximum likelihood estimate of θ is

$$\hat{\theta}_{\mathrm{MLE}} = \mathrm{argmax}[\mathcal{L}(\theta)]\,. \tag{300}$$

In quantum state tomography, the statistical model is Born's rule $[p(E_i) = \mathrm{Tr}(E_i\rho)]$, and its parameters are the matrix elements of ρ. The observed data can be described very simply by a set of POVM effects $\{E_i\}$ and the number of times each effect has been observed, $\{n_i\}$. The likelihood function is

$$\mathcal{L}(\rho) = \Pr(D_{\mathrm{observed}}|\rho)\,, \tag{301}$$

$$= \prod_i \mathrm{Tr}(E_i\rho)^{n_i}\,. \tag{302}$$

The maximum likelihood estimate, $\hat{\rho}_{\mathrm{MLE}}$, is simply the density matrix ρ that maximizes $\mathcal{L}(\rho)$. No general closed-form solutions exist, but finding $\hat{\rho}_{\mathrm{MLE}}$ is a tractable convex optimization problem because the argmax of $\mathcal{L}$ is also the argmax of the log-likelihood function $\log\mathcal{L}(\rho)$,

$$\log\mathcal{L}(\rho) = \sum_i n_i \log(\mathrm{Tr}[E_i\rho])\,, \tag{303}$$

which is concave downward. A variety of numerical algorithms can be used to find the maximum of $\log\mathcal{L}(\rho)$, *constrained* to the convex subset of Hermitian matrices that satisfy (1) $\mathrm{Tr}(\rho) = 1$ and (2) $\rho \geq 0$. The trace constraint is a straightforward linear (holonomic) constraint, but the positivity constraint is trickier. One way to enforce it is by using the nonlinear parameterization

$$\rho = LL^{\dagger}/\mathrm{Tr}(LL^{\dagger})\,, \tag{304}$$

which guarantees both $\rho \geq 0$ and $\mathrm{Tr}(\rho) = 1$. If L is restricted to (complex) lower triangular matrices with real diagonal elements, it is the unique Cholesky factorization of ρ. However, in this parameterization $\log\mathcal{L}(L)$ is not necessarily convex.

In certain circumstances, $\hat{\rho}_{\mathrm{MLE}}$ coincides *exactly* with the linear inversion estimate $\hat{\rho}$ from Eq. (295). If the T matrix from Eq. (295) is invertible (i.e., the tomographic data is informationally complete, but not *overcomplete*), and we ignore the positivity constraint $\rho \geq 0$ and extend the likelihood function to all Hermitian trace-1 ρ for which $\mathcal{L}(\rho) \geq 0$ (including matrices that are not positive semidefinite), then $\mathcal{L}(\rho)$ achieves its maximum uniquely at the linear inversion $\hat{\rho}$. It follows that *if* $\hat{\rho} \geq 0$, then it is the MLE. This can provide significant time savings if/when $\hat{\rho} \geq 0$, because computing Eq. (295) is often much faster than finding $\hat{\rho}_{\mathrm{MLE}}$ numerically.

If $\hat{\rho}$ is not positive, then it follows that $\hat{\rho}_{\mathrm{MLE}}$ must lie on the boundary of the set defined by $\rho \geq 0$—i.e., it will have at least one zero eigenvalue [246]. In this case, it is sometimes useful to *approximate* $\hat{\rho}_{\mathrm{MLE}}$ efficiently by observing that because the linear inversion $\hat{\rho}$ is the maximum of the unconstrained likelihood, $\log\mathcal{L}$ is necessarily quadratic in a neighborhood of $\hat{\rho}$. If the Hessian of $\log\mathcal{L}$ around $\hat{\rho}$ can be efficiently computed, then constrained weighted least squares optimization (instead of generic convex optimization) algorithms can be used to find the $\rho \geq 0$ that maximizes the quadratic approximation to $\log\mathcal{L}$. This approach is sometimes (confusingly) described as "maximum likelihood estimation" in the literature; in fact, it is only an approximation to MLE.

2. Bayesian tomography

Other statistical inference approaches can be applied to tomography. MLE is merely the most common. The most well-studied alternative is *Bayesian* estimation [246–248]. Bayesian statistical inference [249,250] differs from *frequentist* approaches like MLE by assuming (or asserting) that ignorance about an unknown parameter—e.g., a

quantum state—can and should be described by a *probability distribution* over it. Therefore, the standard Bayesian approach to estimating ρ starts by assigning a *prior probability distribution* ("prior") $\mu_{\text{prior}}(\rho)$ to the unknown parameter. The observed data D are used to perform *Bayesian update* on the prior to get a *posterior* distribution by applying *Bayes' rule*:

$$\mu_{\text{post}}(\rho) = \frac{\Pr(D|\rho)\mu_{\text{prior}}(\rho)}{\Pr(D)} = \frac{\mathcal{L}(\rho)\mu_{\text{prior}}(\rho)}{\int_\rho \mathcal{L}(\rho)\mu_{\text{prior}}(\rho)}, \quad (305)$$

where $\mathcal{L}(\rho) = \Pr(D|\rho)$ is the same likelihood function whose maximum defines the MLE. The most common estimator, the *Bayesian mean estimate (BME)* [246], is the mean of the posterior:

$$\hat{\rho}_{\text{BME}} = \int_\rho \rho\mu_{\text{post}}(\rho)\,. \quad (306)$$

Other estimators are possible. However, despite common misconception, the posterior *mode*—i.e., the maximum of μ_{post}—is not a valid estimator. Because μ_{post} is a continuous *measure* over a continuous space, not a function, the posterior probability of any single ρ is precisely zero. For this reason, the posterior cannot have a maximum at any ρ, and *maximum a posteriori (MAP)* estimators do not exist. (MAP estimators exist if and only if the posterior distribution is discrete, but this scenario defines quantum state *discrimination* [251,252] rather than tomography.) Attempts to construct MAP estimators (e.g., Ref. [253]) actually yield a variant of *hedged maximum likelihood* estimation [254].

Bayesian tomography generally requires more computation than MLE, because sampling or integrating a posterior distribution is usually harder than maximizing a convex likelihood function. They also depend critically on the choice of prior, which is a double-edged sword—it is harder to claim "objective" results, but very straightforward to take optimal advantage of "informative" prior information about partially known states. Bayesian estimates are often better behaved and more accurate than frequentist ones [246], and they simplify adaptive tomography [255–257] and uncertainty quantification [258,259]. Software implementations of efficient algorithms [248,257,260] have made Bayesian tomography of states *and* other objects (e.g., gate sets [81,261]) more feasible and accessible.

B. Quantum process tomography

Tomography can also be used to reconstruct (estimate) the CPTP map that best describes a quantum operation (e.g., a logic gate). This is called *quantum process tomography (QPT)* [233,262]. A CPTP map is a linear map on density matrices, a.k.a. a *superoperator* acting on $\mathcal{B}(\mathcal{H})$ (see Sec. II C). In QPT, a CPTP map to be estimated is generally represented either as a *transfer matrix* Λ that acts on a vectorized density matrix $|\rho\rangle\!\rangle$ by matrix multiplication (see Sec. II C 2),

$$|\rho\rangle\!\rangle \mapsto \Lambda|\rho\rangle\!\rangle\,, \quad (307)$$

or as a χ matrix describing

$$\rho \mapsto \sum_{i,j} \chi_{i,j} P_i \rho P_j\,, \quad (308)$$

where $\{P_i\}$ are a basis (often the Pauli basis) for $\mathcal{B}(\mathcal{H})$. The goal of QPT is to estimate a complete mathematical description of the transfer matrix Λ or process matrix χ. Since Λ and χ are equivalent (see Sec. II C), analyses of QPT usually just pick whichever representation is more convenient for the specific protocol being described. We will follow the same convention here.

QPT is performed by choosing an informationally complete set of input states $\{\rho'_j\}$ *and* an informationally complete set of measurements $\{M'_i\}$. The measurements used for QPT must satisfy exactly the same criteria as those used for QST (see Fig. 15), and the input states must collectively span $\mathcal{B}(\mathcal{H})$. Like QST, QPT is very simple in principle. Suppose that $\{\rho'_j\}$ form an informationally complete set of states, so that $\{|\rho'_j\rangle\!\rangle\}$ span $\mathcal{B}(\mathcal{H})$, and $\{E_i\}$ (the union of all the effects of the measurements $\{M'_i\}$) form an informationally complete set of effects, so that $\{\langle\!\langle E_i|\}$ also span $\mathcal{B}(\mathcal{H})$. It follows that the set of superoperators $\{|\rho'_j\rangle\!\rangle\langle\!\langle E_i|\}$ (for all i,j) span the entire space of superoperators. Now, we prepare many copies of every ρ'_j, apply the unknown process to all of them, and then divide the (processed) copies of ρ'_j into groups labeled by i and perform measurement M'_i on the ith group. By doing so, we can estimate every probability

$$p_{i,j} = \Pr(E_i|\rho'_j) = \langle\!\langle E_i|\Lambda|\rho'_j\rangle\!\rangle = \text{Tr}[\Lambda|\rho'_j\rangle\!\rangle\langle\!\langle E_i|]\,. \quad (309)$$

This defines a (large!) set of linear equations that can be solved for Λ in terms of measurable probabilities $p_{i,j}$.

QPT on a system described by a d-dimensional Hilbert space requires at least d^2 distinct input states and enough measurement configurations to perform QST (see Sec. VII A). If only rank-1 PVMs (orthogonal basis measurements) are used, this requires at least $d+1$ distinct measurement configurations, for a total of $d^2(d+1)$ distinct state and measurement configurations, to estimate the $d^4 - d^2$ free parameters of the unknown process. For the special case of n qubits, where $d = 2^n$, this works out to at least $8^n + 4^n$ distinct state and measurement configurations. Achieving this bound requires entangling measurements. If only simultaneous single-qubit measurements are used, then n-qubit process tomography requires 3^n measurement configurations, and thus at least 12^n state and

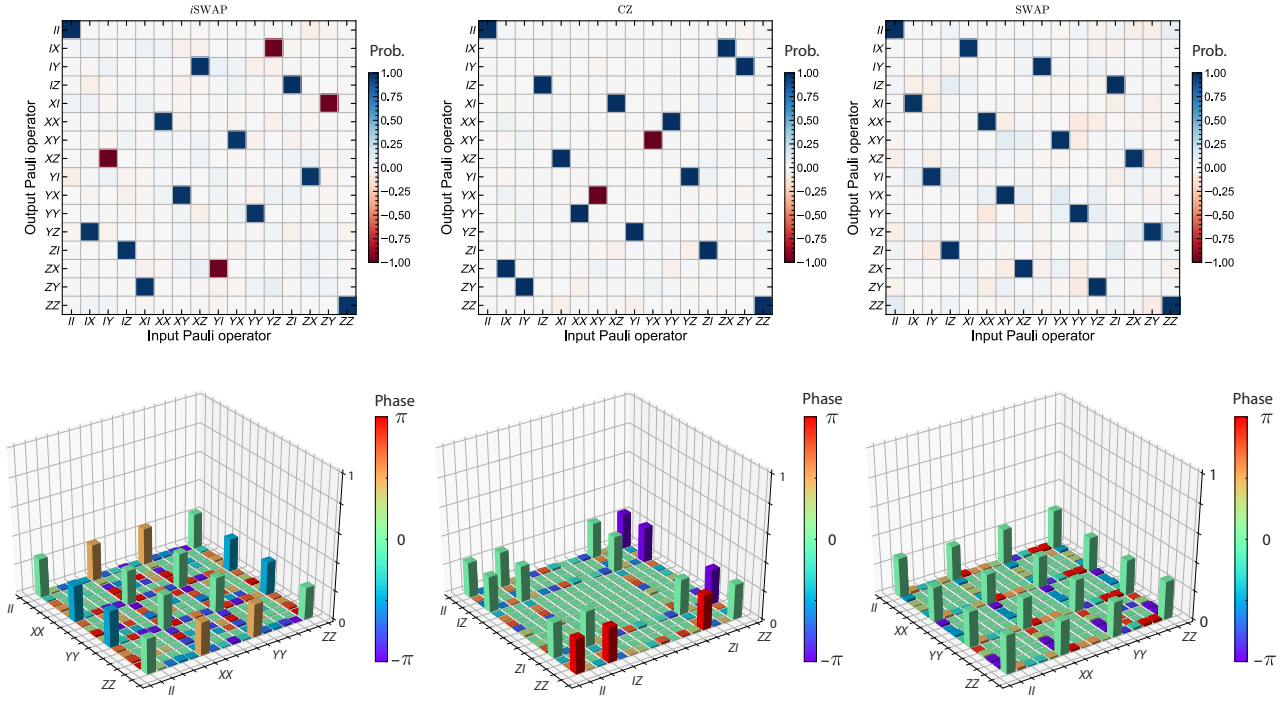


FIG. 17. Quantum process tomography. Top row: PTMs for experimental *i*SWAP, CZ, and SWAP gates reconstructed using QPT [218]. Note that all of the values of a PTM are real and bounded between $[-1, 1]$. Bottom row: χ (process) matrices for the *i*SWAP, CZ, and SWAP gates. The χ matrix is a complex matrix, so each matrix element's magnitude is represented by the height of the corresponding column, and its phase is represented by the column's color. The process fidelities of the gates are 99.32(3)%, 99.72(2)%, and 98.93(5)%, respectively.

measurement configurations. The experimental complexity of QPT grows rapidly for n qubits!

In principle, analysis of QPT data is as simple as "solve Eq. (309) for Λ." But, as the discussion of QST in Sec. VII A illustrates, there are many ways to solve these equations. All of the complications discussed in the context of QST also appear for QPT, which is very nearly isomorphic to QST because of the Choi-Jamiołkowski isomorphism. Much of the QPT literature consists of taking a new QST algorithm (e.g., MLE) and adapting it to QPT. In this tutorial, we do not attempt to explore this literature in detail. Instead, we explain one simple approach to linear-inversion QPT in detail. We assume an n-qubit system, but this analysis generalizes straightforwardly to qudits (see Appendix D 2 b) using a non-Pauli operator basis (see Appendix B 4).

Suppose that by performing QPT experiments we have estimated each of the probabilities in Eq. (309) as

$$\hat{p}_{i,j} \simeq \langle\langle E_i|\Lambda|\rho'_j\rangle\rangle. \tag{310}$$

To solve for Λ, we begin by arranging these probabilities into a matrix P so that $\mathsf{P}_{ij} = \hat{p}_{i,j}$. Next, we construct two matrices by vectorizing the input states and effects in the Pauli operator basis,

$$B = \left(|\rho'_1\rangle\rangle, |\rho'_2\rangle\rangle, \ldots, |\rho'_N\rangle\rangle\right) \tag{311}$$

and

$$A = \begin{pmatrix} \langle\langle E_1| \\ \langle\langle E_2| \\ \vdots \\ \langle\langle E_N| \end{pmatrix}. \tag{312}$$

Choosing the Pauli basis makes the final Λ a Pauli transfer matrix (PTM; see Sec. II C 3). The elements of the input (B) and output (A) matrices are then

$$B_{ij} = \frac{1}{\sqrt{d}}\mathrm{Tr}[P_i\rho_j], \tag{313}$$

$$A_{ij} = \frac{1}{\sqrt{d}}\mathrm{Tr}[E_iP_j]. \tag{314}$$

Now, using A, B, and P, we can express Eq. (309) as

$$\mathsf{P} = A\Lambda B, \tag{315}$$

and extract Λ by inverting the A and B matrices [264,265]:

$$\Lambda = A^{-1}\mathsf{P}B^{-1}. \tag{316}$$

When the input states and/or effects are overcomplete, a Moore-Penrose pseudoinverse can be used instead [266].

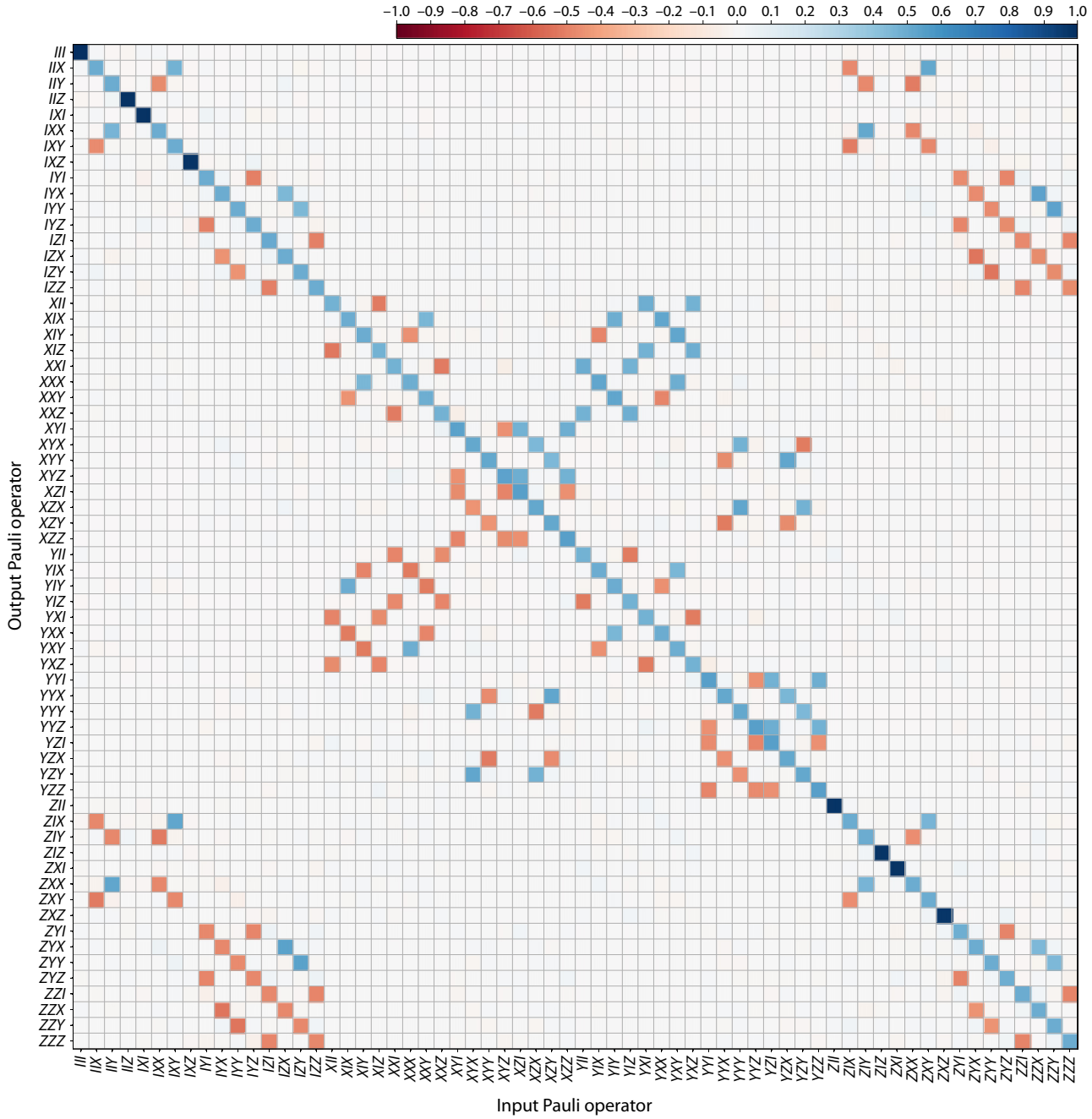


FIG. 18. Three-qubit QPT. PTM of a three-qubit *i*Toffoli gate [263]. A three-qubit PTM contains $16^3 - 4^3 = 4032$ independent parameters, which can be estimated from a minimum of $4^3 \times 3^3 = 1728$ independent experiments. The process fidelity is estimated to be 97.1(8)%.

An equivalent approach, which offers complementary intuitions about QPT, is to use the data to perform quantum state tomography on $\Lambda|\rho_j'\rangle\rangle$ for each input state ρ_j'. Then, if we write both the input and output states as vectors in the Pauli basis, we can simply solve for the PTM Λ using least-squares fitting. Figure 17 shows PTMs and χ matrices reconstructed this way from experimental QPT on two-qubit *i*SWAP, CZ, and SWAP gates implemented on a superconducting quantum processor. A minimal set of input states $\{|0\rangle\langle 0|, |1\rangle\langle 1|, |+\rangle\langle +|, |-\rangle\langle -|\}^{\otimes 2}$ was used.

Adding a third qubit increases the number of experimental configurations required from 144 to 1728, illustrating the rapid growth of QPT's experimental complexity with n. Figure 18 shows the PTM for a three-qubit *i*Toffoli gate, which contains 4032 independent parameters estimated from 1728 distinct experiments.

As we noted at the beginning of Sec. VII, both QST and QPT are vulnerable to SPAM errors, which cause systematic bias in the estimate. Therefore, the process fidelities quoted in Figs. 17 and 18 do not separate gate errors from SPAM errors. For this reason, experimental tomography of gates has moved toward tomographic reconstruction methods that characterize SPAM errors and gate errors simultaneously and self-consistently, such as gate set tomography (Sec. VII D). A variation of QPT that leverages ideas from gate set tomography to eliminate or correct SPAM bias has also been proposed [266].

The examples above illustrated *linear-inversion* QPT. But, just as density matrices reconstructed using linear-inversion QST can easily violate the positivity constraint $\rho \geq 0$, superoperators reconstructed using linear-inversion QPT can violate complete positivity (CP; see discussion in Sec. II C). There are many ways to constrain a reconstructed PTM or process matrix to be CP, including MLE [264,267,268] or projection algorithms [269]. CP-constrained MLE can be done in a variety of ways (e.g., via semidefinite programs [264]), but the easiest approach to understand uses the Choi-Jamiołkowski isomorphism. In this approach, the process is parameterized by its χ matrix, which is isomorphic to a density matrix on a larger system. Now, MLE can be performed using algorithms designed for state tomography (although an additional constraint on the χ matrix, corresponding to trace preservation, must be added).

C. Quantum measurement tomography

The goal of *quantum measurement tomography (QMT)* is to estimate the parameters of a mathematical model for a quantum measurement operation. For the most common case where the measurement is modeled by a POVM $M = \{E_i\}$ (see Sec. II B 2), QMT requires applying the unknown measurement to an informationally complete set of input states $\{\rho'_j\}$ (see Fig. 15). So, QMT can be seen as the complement to QST (Sec. VII A): where unknown states are estimated by performing a range of known measurements on them, an unknown measurement is estimated by applying it to a range of known input states.

A POVM describing an m-outcome measurement on a d-dimensional quantum system comprises m $d \times d$ effects $E_i \geq 0$. Estimating such a POVM requires repeatedly applying it to an informationally complete set of at least d^2 input states whose density matrices span $\mathcal{B}(\mathcal{H})$. For an n-qubit system, 4^n linearly independent input states are required.

The most common QMT procedure is to choose 4^n linearly independent input states from the 6^n tensor products of single-qubit Pauli eigenstates, prepare N samples of each, apply the unknown measurement to all the samples, record the outcome statistics, and estimate the probabilities

$$p(i|j) = \Pr(E_i|\rho'_j) = \mathrm{Tr}[E_i\rho'_j] = \langle\langle E_i \,|\, \rho'_j \rangle\rangle . \tag{317}$$

There is no uniquely good way to select a subset of 4^n Pauli eigenstates. Optimal accuracy is achieved when the input states form a 2-design, but even for a single qubit, achieving this optimum requires either (1) choosing non-Pauli eigenstates such as a SIC-POVM [185], or (2) using all six Pauli eigenstates.

Analysis of QMT data proceeds identically to QST and QMT. Equation (317) is solved using the same techniques and methods—e.g., linear inversion, least-squares, or MLE—to find each E_i, and thus the entire unknown POVM M. Measurement tomography implies slightly different constraints than state or process tomography; each effect E_i must be positive semidefinite, but the analog of the trace or TP constraints is that the *sum* $\sum_i E_i$ must equal $\mathbb{I}$. This requires technical changes to constrained MLE algorithms, but no conceptual novelty [270,271].

Tomographic estimation of mid-circuit measurements requires a different model. POVMs model *terminating* measurements, and mid-circuit measurements are modeled not by POVMs, but by *quantum instruments* [Eq. (116)]. Tomographic reconstruction of quantum instruments is a reasonably straightforward fusion of QPT and POVM tomography, and the interested reader is referred to Refs. [177,189,190].

Sometimes, it is adequate or desirable to use a simpler model for terminating measurements instead of a full POVM. A *response* (or *confusion*) matrix provides such a model when the unknown measurement is intended to measure a canonical basis (e.g., the computational basis). The response matrix R is a $d \times d$ matrix whose elements $R_{ij} = p(E_i|\rho_j)$ are the probabilities of getting outcome $E_i \approx |i\rangle\langle i|$ if state $\rho_j \approx |j\rangle\langle j|$ is prepared and measured. For example, a single-qubit response matrix is given by

$$R = \begin{pmatrix} p(0|0) & p(0|1) \\ p(1|0) & p(1|1) \end{pmatrix} . \tag{318}$$

R can be estimated in the obvious way, by preparing each basis state as accurately as possible, applying the measurement, and recording the empirical probabilities of each outcome, as shown in Fig. 19. Response matrix estimation requires preparing only d input states, whereas QMT estimation of a POVM requires d^2. It provides a subset of the information in a POVM. For example, it can be used to compute the readout fidelity defined in Eq. (254), which is given by $\mathrm{Tr}(R)$, but it does not predict the results of performing the measurement on superpositions of the canonical basis states. Like QST and QPT, QMT's experimental complexity grows exponentially with the number of qubits n, making it effectively infeasible for more than a few qubits. However, measurements can be characterized much more efficiently if a valid ansatz applies.

A common and powerful ansatz is to assume that crosstalk in multiqubit readout is negligible. If correlated

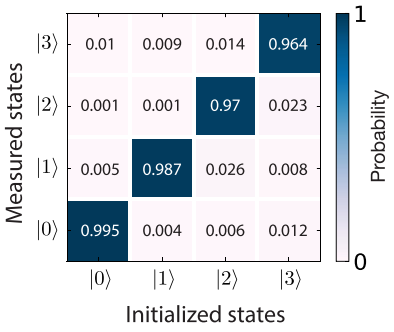


FIG. 19. Response matrix. Heralding and measuring the states of a qudit of dimension $D = 4$ yields the probabilities $p(i|j)$ of measuring state $|i\rangle$ after preparing state $|j\rangle$. These constitute the response matrix R. (The data are reproduced with permission from Ref. [272].)

readout errors are negligible, then a multiqubit response matrix R or POVM M can be approximated as the tensor product of each individual qubit's R or M [273]. This holds true for mid-circuit measurements modeled by quantum instruments as well. Crosstalk-free models of measurements can generally be characterized using a set of input states whose size scales just *linearly* with the number of qubits. But because crosstalk effects are often nonnegligible [177,190,274], crosstalk-free models should be viewed with caution as a potentially useful approximation. Full POVM characterization is usually necessary for accurate assessment and modeling. Some forms of crosstalk (and other errors in measurement) can be effectively eliminated by "twirling" readout error into stochastic bit-flip channels [182,183]. Twirled measurements display less crosstalk, and are almost always more consistent with the response matrix model, than native measurements.

D. Gate set tomography

State, process, and measurement tomography are powerful tools for diagnosing errors in a quantum processor. However, each of these protocols implicitly assumes the existence of a precalibrated *reference frame* (see Fig. 15) of perfect states and/or measurements. Errors in the operations that define such a reference frame can bias the tomographic reconstructions, and lead to incorrect models for the operations under test. *Gate set tomography (GST)* [156,275] is a family of calibration-free approaches to tomography that explicitly acknowledge that all elements of a quantum computer's *gate set*—the native state preparations, measurements, and logic gates (see Sec. II E)—are subject to errors. GST protocols are able to reconstruct self-consistent mathematical representations of a quantum computer's native gate set and the errors afflicting it.

Around 2012, groups at IBM [194] and Sandia National Laboratories [195] independently identified the need for calibration-free characterizations of quantum operations. IBM approached this problem using a so-called "overkill" tomography protocol that utilizes all circuits of depth 3 or less and fits a gate set model with MLE. Sandia's "linear GST" method uses similar circuits to standard QPT and fits a model with linear inversion. Variants of these early protocols are still in use to some extent, but since their introduction the family of GST protocols has evolved significantly. It now encompasses a rather broad set of experiment design and data analysis techniques for self-consistently estimating the parameters of a gate set model (see Sec. II E). In this tutorial, we limit our discussion to two essential protocols: *linear GST*, mentioned above, and *long-sequence GST*, which uses long, structured quantum circuits and iterative MLE. Significant extensions to these protocols [80,276] have introduced approaches for characterizing larger processors and devices with mid-circuit measurements [55], and other work [82] introduced an efficient Kalman filter formalism for GST that replaces the MLE analysis with a Bayesian estimator capable of streaming real-time data processing. Experimental implementations of GST can be found in many papers, including (but not limited to) Refs. [46,78,79,167,195,277,278].

We use the term "GST" without qualification to indicate "long-sequence GST." Reference implementations of linear and long sequence GST can be found in the `pyGSTi` python package [279]. Throughout the rest of this subsection, we use the term "gate set" to describe both the ensemble of logical instructions available on a given quantum computer (e.g., "prepare $|0\rangle$," "Hadamard gate on qubit 3," "measure qubit 1," etc.), and the mathematical representations of those objects (e.g., density matrices, transfer or process matrices, and POVM elements). In discussing those mathematical objects, we follow the conventions of Eq. (319a) for defining a gate set $\mathcal{G}$:

$$\mathcal{G} = \left\{ \left\{ |\rho^{(i)}\rangle\!\rangle \right\}_{i=1}^{N_\rho}, \{G_i\}_{i=1}^{N_\mathrm{G}}, \left\{ \langle\!\langle E_i^{(m)}| \right\}_{m=1,i=1}^{N_\mathrm{M},N_\mathrm{E}^{(m)}} \right\}. \quad (127)$$

Here, N_ρ is the number of native state preparations, N_G is the number of native gates, N_M is the number of native measurements, and $N_E^{(m)}$ is the number of outcomes for the m^{th} native measurement. In most cases, $N_\rho = N_M = 1$.

Gate set models often have many parameters, and their size grows very rapidly with the number of qubits n. Each operation in the gate set is a matrix whose dimension grows exponentially with n. Moreover, the total number of possible n-qubit operations can grow *combinatorially*. Fitting a model with exponentially many parameters requires experimental complexity (and time) that also grows exponentially with n. For these reasons, standard GST protocols have so far been applied only to one- and two-qubit systems.

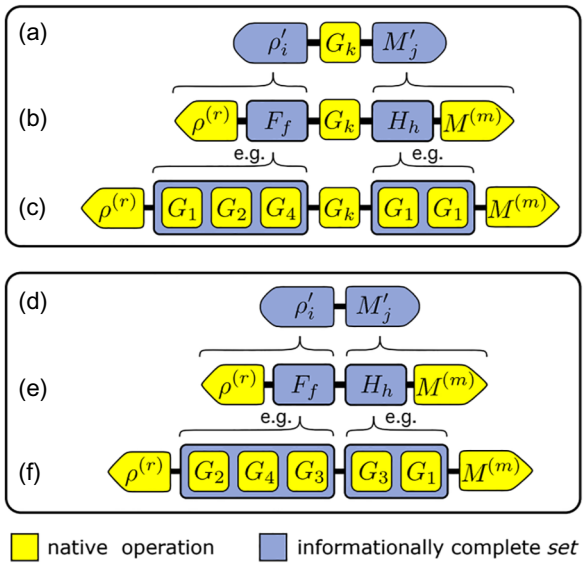


FIG. 20. Linear gate set tomography. Structures of the two types of circuits required by the LGST algorithm. **Upper panel:** each native gate, G_k, is sandwiched between the elements of informationally complete sets of *effective state preparations*, $\{\rho_i'\}$, and of *effective measurements,* $\{M_j'\}$. These are the same circuits that QPT requires to characterize G_k. (a) Shows these circuits in their simplest form, with each informationally complete set displayed as a unit. (b) Depicts the common case when the set of effective preparations (measurements) is implemented by following (preceding) a single native state preparation (measurement) operation with a *fiducial circuit* F_f (H_h). (c) Exemplifies that the fiducial circuits are composed of native gates G_i and gives the circuit entirely in terms of native operations. **Lower panel:** because LGST does not assume knowledge of ρ_i' and M_j', it requires circuits that sandwich nothing between pairs of fiducials in order to be self-calibrating. The circuit diagrams in (d)–(f) parallel those in (a)–(c). LGST also requires the circuits that perform state (measurement) tomography on ρ (M), but these are not explicitly shown. They are similar to (d)–(f) (replacing ρ' with ρ or M' with M), and are actually included as a subset of these circuits when the gate set contains only a single native state preparation (measurement) and one of the preparation (measurement) fiducial circuits is the empty (do-nothing) circuit. (Figure and caption reproduced with permission from Ref. [156].)

1. Linear GST

Linear gate set tomography (LGST) is a self-consistent approach to simultaneous state, process, and measurement tomography that uses short quantum circuits and reconstructs a gate set model using linear inversion (see Fig. 20). Like QPT, it assembles elements of the gate set to construct informationally complete (see Sec. V A) sets of states and measurements, which are then used to probe the errors in elementary logic operations. Most quantum computing systems can natively prepare only a single initial state (typically $|000\cdots\rangle$) and perform measurements only in a single basis (e.g., the computational basis). So, a full, informationally complete set of states and measurements must be constructed from these native operations by the application of short *fiducial* gate sequences. For instance, measurement in the $\{|+\rangle, |-\rangle\}$ basis can be performed by preceding a computational basis measurement by a Hadamard operation. Informationally *over*complete fiducial sets are often used in the literature, but for simplicity of presentation, we assume exact informational completeness. See the Appendix of Ref. [156] for the general case.

Given an informationally complete set of fiducial states $\left\{|\rho_j'\rangle\rangle\right\}_{j=1}^{N_\rho}$ and an informationally complete set of fiducial measurement effects $\left\{\langle\langle E_i'|\right\}_{i=1}^{N_E}$, the LGST protocol prescribes a set of circuits whose output distributions provide sufficient information to estimate the parameters of a gate set model. To see how the protocol works, it is convenient to start by collecting all the fiducial states and measurement effects into matrices A and B defined as

$$A = \begin{pmatrix} \langle\langle E_1'| \\ \langle\langle E_2'| \\ \vdots \\ \langle\langle E_N'| \end{pmatrix} \tag{319}$$

and

$$B = \left(|\rho_1'\rangle\rangle, |\rho_2'\rangle\rangle, \ldots, |\rho_{N_\rho}'\rangle\rangle\right) . \tag{320}$$

Now, the parameters of a gate G_k can be estimated by first estimating a matrix of probabilities P_k defined componentwise as

$$[\mathsf{P}_k]_{i,j} = \langle\langle E_i'| G_k |\rho_j'\rangle\rangle , \tag{321}$$

or as a matrix equation,

$$\mathsf{P}_k = AG_kB . \tag{322}$$

Measuring P_k is essentially standard QPT, but in the absence of a precalibrated reference frame, neither A nor B is known, so we cannot solve Eq. (322) directly for G_k.

Instead, LGST also measures an additional set of circuits that would correspond to QPT on the null operation. Their probabilities, which can be estimated from the circuits' observed outcome statistics, form a *Gram matrix* $\tilde{\mathbb{1}}$ whose elements are

$$[\tilde{\mathbb{1}}]_{i,j} = \langle\langle E_i'|\rho_j'\rangle\rangle . \tag{323}$$

The Gram matrix is precisely equal to

$$\tilde{\mathbb{1}} = AB . \tag{324}$$

Since the fiducial states and measurements are (by assumption) informationally complete, both A and B are square

invertible matrices, and $\tilde{\mathbb{I}}^{-1} = B^{-1}A^{-1}$. It follows that multiplying both sides of Eq. (323) by $\tilde{\mathbb{I}}^{-1}$ yields

$$\tilde{\mathbb{I}}^{-1}\mathsf{P}_k = B^{-1}G_kB\,, \tag{325}$$

or, solving for G_k,

$$G_k = B\tilde{\mathbb{I}}^{-1}\mathsf{P}_kB^{-1}\,. \tag{326}$$

This equation defines the process matrices for every gate in the gate set, in terms of physically measurable quantities $\tilde{\mathbb{I}}$ and P_k, *up to an unknown superoperator* B. It turns out that B is not just unknown, it is unknowable—B embodies the *gauge freedom* in gate set models. So Eq. (327) defines a complete estimate of all gates in the gate set up to gauge freedom.

The native state preparations $\{\rho^{(i)}\}_{i=1}^{N_\rho}$ and measurements $\left\{E_j^{(m)}\right\}_{m=1,j=1}^{N_M,N_E^{(m)}}$ can be estimated (in the same gauge) by constructing the following vectors:

$$[\mathbf{R}^{(l)}]_j = \langle\langle E'_j|\rho^{(l)}\rangle\rangle\,, \tag{327}$$

$$[\mathbf{Q}_l^{(m)}]_j = \langle\langle E_l^{(m)}|\rho'_j\rangle\rangle\,, \tag{328}$$

using empirically measured circuit outcome probabilities. We write them as

$$\mathbf{R}^{(l)} = A\,|\rho^{(l)}\rangle\rangle\,, \tag{329}$$

$$\mathbf{Q}_l^{(m)T} = \langle\langle E_l^{(m)}|\,B\,, \tag{330}$$

and use the Gram matrix identity $\tilde{\mathbb{I}} = AB$ to write all the elements of a gate set model in terms of measurable quantities and a gauge transformation B as

$$G_k = B\tilde{\mathbb{I}}^{-1}\mathsf{P}_kB^{-1}\,, \tag{331}$$

$$|\rho^{(l)}\rangle\rangle = B\tilde{\mathbb{I}}^{-1}\mathbf{R}^{(l)}\,, \tag{332}$$

$$\langle\langle E_l^{(m)}| = \mathbf{Q}_l^{(m)T}B^{-1}\,. \tag{333}$$

Any invertible matrix B is a valid choice, and the predicted outcome probabilities of circuits are entirely independent of B. Different choices of B simply define different gauges (bases) in which the gate set can be written. Thus, no physical experiment can single out a "proper" gauge. Inconveniently, most metrics of performance (Sec. IV) are *not* gauge-invariant (independent of B). Therefore, it is customary and necessary to choose a convenient gauge in which to report the estimated $\hat{\mathcal{G}}$. This is discussed in Sec. II E 2; an easy starting choice is to use the B defined by the *intended* fiducial states.

In the above discussion, P_k is a matrix of circuit outcome probabilities that must be estimated from data. The maximum likelihood estimator for those probabilities is simply the observed frequency. If the circuit is run N times, then finite sample fluctuations will lead to error in the estimate that scales like $\mathcal{O}(1/\sqrt{N})$. Since G_k is linear in P_k [Eq. (332)], the error bars on the estimate of the G_k transfer matrix also scale as $1/\sqrt{N}$. This is the so-called "standard quantum limit" for parameter estimation [280], and it results here from the fact that each gate is only used once per circuit (excluding any potential uses in creating the fiducials). Long-sequence GST evades the standard quantum limit by using long circuits that amplify errors and achieve "Heisenberg" scaling in estimation precision.

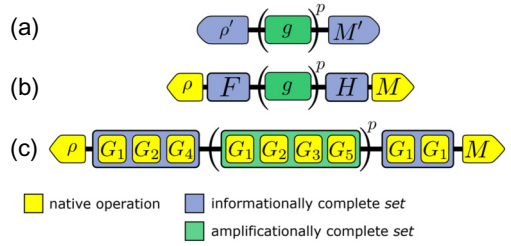


FIG. 21. Long-sequence gate set tomography. The structure of circuits in the standard LSGST experiment design, shown in increasing detail. (a) Each GST circuit consists of an effective state preparation ρ' followed by a *germ* circuit g repeated p times, followed by an effective measurement $M' = \{E'_i\}$. (b) Effective preparations are usually implemented by a native state preparation ρ followed by a *preparation fiducial circuit* F, and effective measurements are usually implemented by a *measurement fiducial circuit* H followed by a native measurement M. (c) Writing the fiducials and germ in terms of the native gate operations reveals how the native operations of a gate set compose to form a GST circuit. (Figure and caption reproduced with permission from Ref. [156].)

2. Long-sequence GST

Long-sequence gate set tomography (LSGST) is an approach to self-consistent tomography of quantum gate sets that can beat the standard quantum limit [197]. It achieves this by adding two innovations beyond LGST: (i) longer quantum circuits that amplify gate errors, and (ii) a new statistical estimation protocol to analyze long-circuit data.

LSGST circuits are formed in a similar fashion to LGST circuits. First, one selects informationally complete sets of state preparation and measurement fiducials, as in LGST. Where LGST uses these fiducials to probe each gate G_k in the gate set, long-sequence GST uses them to probe each of an *amplificationally complete* list of "germs." A germ is a short sequence of native gates, chosen so as to amplify certain errors (see below). The circuits run by LSGST comprise a fiducial state preparation, an p-fold repeated germ, and a fiducial measurement (see Fig. 21). The "germ powers" p are typically chosen to make the length of the p-fold repeated germ as close as possible to (but not greater than) a logarithmically spaced integer, $1, 2, 4, 8, \ldots, p_{\max}$.

These many-fold repeated germs are what enable LSGST to achieve Heisenberg-limited scaling.

To see how this works, consider repeating a single gate G many (p) times. The resulting process can be computed from a spectral decomposition of the gate,

$$G \doteq R \begin{pmatrix} e^{\phi_1} & & \\ & e^{\phi_2} & \\ & & \ddots \end{pmatrix} R^{-1}, \tag{334}$$

$$G^p \doteq R \begin{pmatrix} e^{p\phi_1} & & \\ & e^{p\phi_2} & \\ & & \ddots \end{pmatrix} R^{-1}, \tag{335}$$

where $\mathrm{diag}(e^{\phi_1}, e^{\phi_2}, \ldots)$ is a diagonal matrix of (generally complex) eigenvalues and R is the change-of-basis matrix between the original and diagonalized bases. As we saw above, probing G with an informationally complete set of fiducials allows us to estimate its eigenvalues, and thus the complex phases ϕ_i, with uncertainty $\mathcal{O}(1/\sqrt{N})$. Similarly, probing G^p allows us to estimate the *amplified* phases $p\phi_i$ with uncertainty $\mathcal{O}(1/\sqrt{N})$, giving a $\mathcal{O}(1/p\sqrt{N})$ uncertainty in the estimate of ϕ_i. This example also helps explain why GST does not exclusively use the longest sequences: knowledge of $e^{p\phi}$ for $p \gg 1$ is generally insufficient to deduce ϕ, because the logarithm is multivalued. Data from shorter sequences are necessary to determine the correct branch.

Repeating G many times *amplifies* its eigenvalues, allowing them to be estimated to high precision. But it does not increase sensitivity to *axis errors*, so it does not enable high-precision estimation of the matrix of eigenvectors R. Amplifying the misalignment of the axis between the gates requires choosing and repeating composite germs that consist of products of native gates. Each germ defines a quantum operation, which can be described by a transfer matrix. Errors in a germ's component gates propagate to become errors in the germ. Some of them change the eigenvalues of the germ's transfer matrix. Those errors—often linear combinations of errors on different gates—are precisely the ones that the germ amplifies, and which can be measured to high precision by repeating that germ. A set of germs is called *amplificationally complete* for a gate set $\mathcal{G}$ if knowledge of all the germs' eigenvalues is sufficient to reconstruct all of the nongauge, non-SPAM degrees of freedom in the gate set.

To analyze what errors a given germ amplifies, we assume that errors act as small perturbations to the target operation of a gate. In that limit, a perturbation to the germ's target operation is amplified by the germ if and only if it commutes with the target operation. It follows that each germ amplifies a subspace of perturbations to the gate set. A complete LSGST experiment design can be constructed by searching over possible germs (generally starting with the shortest) until the amplified directions span the nongauge subspace; see Ref. [156] and the discussion of Sec. V A.

Fitting LSGST data is typically done using iterative maximum likelihood or least-squares optimization. This approach begins by fitting a model to the shortest ($p = 1$) sequences, and then using that to seed the optimizer for the next round, which includes the $p = 1$ and $p = 2$ circuits. This procedure repeats until all circuits have been added and the optimizer has converged. This approach helps to avoid the wrong branch issue described above, and in practice is extremely robust.

LSGST experiments designed this way are very overcomplete. For example, constructing a two-qubit LSGST experiment for a relatively simple gate set this way can easily prescribe more than 30 000 circuits! Many of these circuits provide redundant information, and can be removed from the experiment design with minimal loss of estimation accuracy. Reference [204] discusses techniques for constructing efficient, near-minimal GST experiments. In practice, these techniques have reduced the experimental complexity of two-qubit GST by more than an order of magnitude, making it feasible and competitive with randomized benchmarking.

3. (In)validation of gate set models

Many useful performance metrics can be extracted from a high-precision estimate of a gate set model, including both gauge-dependent and gauge-independent metrics discussed in Sec. II E 2. The *error generator* framework discussed in Appendix A can be used to identify observed errors with physical mechanisms. These analyses help quantify and classify the "in model" errors captured by the gate set model.

But, in practice, quantum computing systems often experience "out of model" errors that violate the assumptions of the gate set model. Such errors are often termed *non-Markovian* because gate set models are designed to capture arbitrary Markovian errors [197], and thus (arguably) define what it means for errors to be Markovian, as discussed in Sec. III G. Examples of non-Markovian errors include low-frequency drift, leakage, and heating of auxiliary degrees of freedom (e.g., the trapped-ion motional mode used in Mølmer-Sørenson gates).

When non-Markovian effects meaningfully impact circuit outcome statistics, it is very unlikely that any Markovian gate set model will be statistically consistent with the observations. This is an interesting difference between GST and QPT; because QPT experiments use the unknown process just once, and are usually not very overcomplete, they do not generally distinguish between Markovian and non-Markovian effects. In contrast, LSGST experiment

designs use the unknown gate[s] many times (which amplifies non-Markovian effects) and are typically overcomplete. This overcompleteness can be leveraged to quantify, using statistical tests, how well or poorly an estimated gate set model fits the observed data. Model violation analysis provides insight into the presence, size, and sometimes nature of non-Markovian errors in GST experiments.

The primary tools for such a "goodness-of-fit" analysis are the *log-likelihood ratio test* and *Wilks' theorem* [281]. They make extensive use of the log-likelihood ratio statistic $\lambda \equiv 2(\log \mathcal{L}_{\max} - \log \mathcal{L})$ that compares the likelihood $\mathcal{L}$ of an estimated GST model to the likelihood $\mathcal{L}_{\max}$ of a much larger *maximal* (a.k.a. *saturated*) model that is guaranteed to fit the data well. The maximal model commonly used in GST treats each circuit as a totally independent experiment, assigning it an independent multinomial outcome distribution whose parameters are fit directly to that circuit's observed frequencies. Wilks' theorem [281] states that if the GST model is valid, then this log-likelihood ratio will be a χ^2_k random variable,

$$2(\log \mathcal{L}_{\max} - \log \mathcal{L}) \sim \chi^2_k\,, \tag{336}$$

where k is the difference between the number of parameters in the maximal model, $N_{\max}$, and the number of *nongauge* parameters in the estimate, N_{nongauge}: $k = N_{\max} - N_{\text{nongauge}}$. A simple way to quantify the significance of the observed model violation is to compute and report the number of standard deviations by which the log-likelihood ratio exceeds its expected value under a χ^2_k hypothesis:

$$N_\sigma \equiv \frac{2(\log \mathcal{L}_{\max} - \log \mathcal{L}) - k}{2\sqrt{k}}\,. \tag{337}$$

If $N_\sigma \approx 1$, then the estimated gate set model fits the data as well as possible, suggesting that the device's errors are mostly Markovian. However, $N_\sigma \gg 1$ indicates convincing statistical evidence for the presence of non-Markovian dynamics.

N_σ is a statistical measure of model violation, and captures the strength of the observed evidence for non-Markovian dynamics. What it does *not* do is quantify *how much* non-Markovianity is present in any physically meaningful units. It is not a measure of *effect size*. For example, simply doubling the number of shots performed for each circuit will (on average) double the log-likelihood ratio statistic and thus N_σ. The log-likelihood ratio statistic scales linearly with the amount of *data* available.

It is also useful to have a measure of the effect size of non-Markovian errors. In principle this could be extracted from a (much) larger model that is able to capture any expected non-Markovian effects. But such models do not yet exist (although methods do exist for low-frequency noise [96]), and doing so would require designing an experiment that is sensitive to all of the parameters and fitting it to data. *Wildcard models* [217] provide an alternative. Wildcard models weaken the predictions of statistical error models just enough that they become consistent with observed data, and the amount of weakening required to do so constitutes a measure of model violation effect size. The parameters of a wildcard model can, with care, be interpreted as measuring how much non-Markovian error is present in the data. The wildcard error can then be compared to various error metrics, such as diamond distance (see Sec. IV C 1), to determine whether or not the GST model is trustworthy. A full discussion of wildcard models is out of scope for this tutorial, but the interested reader is encouraged to consult Ref. [217]. References [78] and [79] provide examples of how this type of analysis can be used in practice.

VIII. RANDOMIZED BENCHMARKS

Randomized benchmarking (RB) protocols are a broad suite of methods that use varied-depth random circuits to quantify the rates of errors in a gate set (see Secs. II E and IV E). RB was initially developed in the mid- to late-2000s [160,282,283] to circumvent two of main the limitations of quantum process tomography (QPT; see Sec. VII B); QPT is corrupted by SPAM errors, and it is inefficient in the number of qubits (n). There are now dozens of distinct RB protocols, each with their own purposes, strengths, and limitations. In this section, we review many of the most widely used RB methods. In the first half of this section, we discuss the RB protocols that estimate a *single* error rate for a set of gates:

(a) *Standard RB* (Sec. VIII B). This is the *de facto* standard RB protocol, which is typically used to benchmark gates that implement the one- or two-qubit Clifford group.
(b) *Native gate RB protocols* (Sec. VIII C). These are a family of protocols that can directly benchmark a system's native gates, instead of using those gates to create all the Clifford group elements (as in standard RB). Protocols within this family include *direct RB*, *binary RB*, *mirror RB*, and *cross-entropy benchmarking*.
(c) *RB for general groups* (Sec. VIII D). This is a family of protocols for benchmarking sets of gates that form groups that are not unitary 2-designs. The most prominent such method is *character RB*.

There are a variety of RB protocols that measure quantities that are more complex or fine-grained than just a single error rate for a gate set (e.g., individual gate error rates). Many of these methods are adaptations of the foundational RB protocols presented in Secs. VIII B–VIII D. We discuss the following:

TABLE IV. Comparison of randomized benchmarks. Here, we summarize the primary randomized benchmarks discussed in this tutorial (note that this tutorial does not cover every extant RB method). Randomized benchmarks do not all measure the same quantity, and so the first consideration when selecting an RB method is to consider what it measures (second column). The scalability of randomized benchmark is also often an important consideration when selecting an RB method, and here we give each method an *informal* scalability score (between 0 and 10) that approximately summarizes how scalable it is. The scalability score is subjective, and we intend it only as very rough guide. For example, mirror RB, binary RB, and cycle benchmarking are all given a score of 8/10 for scaling because their classical computations required are simple and fast for any number of qubits (n), but they require both implementing an n-qubit layer/cycle of native gates and measuring all n qubits with significantly nonzero fidelity. Key advantages and disadvantages of each method are also summarized.

Method	What it Measures	Scalability Score	Advantage(s)	Disadvantage(s)
Clifford RB	Average error per Clifford gate (r)	2	*de facto* standard protocol	Poor scaling; does not reliably measure error rate of native gates
Direct RB	Weighted-average error per gate in user-chosen gate set (r_Ω)	3	Can measure error per native gate	Scales better than standard RB but still limited, especially for non-Clifford gates
Binary RB	Weighted-average error per gate in user-chosen gate set (r_Ω)	8	Both reliable and scalable	All gates must be Clifford
Mirror RB	Weighted-average error per gate in user-chosen gate set (r_Ω)	8	Scalable, including for non-Clifford gates	Small but systematic underestimate of gate error
Cross-Entropy Benchmarking	Weighted-average error per gate in user-chosen gate set (r_Ω)	6	Native gate does not need to be a Clifford	Need to classically simulate circuits, so very expensive beyond $n \sim 20$ qubits, and infeasible beyond $n \sim 60$
Character RB	Average error per group-element gate	Depends on gate set used	Can benchmark gate sets that are not and do not generate unitary 2-designs	Sample inefficient (requires lots of data) for some groups
Simultaneous RB	Average error per simultaneous Clifford gate	9	Captures effects of crosstalk on qubits	Simple for one-qubit gates but complex scheduling for $n > 2$ qubit gates
Interleaved RB	Average error per dressed native Clifford gate	1	Simple and widely used	Unreliable; native gate must be a Clifford
Cycle Benchmarking	Average error per dressed Clifford cycle/layer	8	Can estimate the error rate of dressed n-qubit cycle for large n	Cycle must be Clifford (in most cases); need to measure multiple decays
eXtended RB	Average unitarity and stochastic error per Clifford gate	1	Able to separate coherent from incoherent contributions to the AGSI	Inefficient (does not amplify coherent errors); not scalable (uses state tomography)
Speckle Purity Benchmarking	Stochastic error per layer of native gates	3	Enables distinguishing coherent and stochastic contributions to XEB error rate	Similar disadvantages to XEB
Iterative RB	Coherent error per interleaved Clifford gate	1	Conceptual simplicity, amplifies coherent errors	Only works for over/under-rotation errors
Leakage RB	Leakage rate per Clifford gate	2	Makes RB more reliable in the presence of leakage, measures leakage rate(s)	Data may not follow an exponential due to no randomization in the leakage space

(a) *Simultaneous RB* (Sec. VIII E). Simultaneous RB is a simple and widely used technique for measuring the impact of simultaneous gate operations across multiple qubits. It can be used to quantify crosstalk errors between qubits.
(b) *Interleaved RB* (Sec. VIII F). Interleaved RB is a technique for estimating the infidelity of individual gates, but it has important limitations.
(c) *Cycle benchmarking* (Sec. VIII G). Cycle benchmarking is a scalable method for estimating the infidelity of layers of gates.
(d) *Purity benchmarking protocols* (Sec. VIII H). These are a family of protocols for estimating how much of a gate set's error is due to coherent and incoherent errors.
(e) *RB protocols for non-Markovian errors* (Sec. VIII I). These are a family of protocols for estimating the rates of various kinds of non-Markovian errors, such as leakage.

We begin this section with some mathematical background that is important for understanding and describing the various randomized benchmarks that we outline above.

A. Mathematical preliminaries

Despite being rather simple to implement, the mathematical theory of RB protocols is surprisingly deep and elegant. Describing it in full detail is well beyond the scope of this tutorial. But several of the most important concepts from this theory are found commonly even in the experimental literature. In this subsection, we introduce those few mathematical concepts that are most helpful for reading and understanding papers on RB and related benchmarking protocols. These topics include the following:

(a) twirling over a group,
(b) Schur's lemma, and
(c) unitary 2-designs.

The pragmatic reader can skip to Sec. VIII B, where the RB protocol discussions begin.

A number of QCVV techniques, including RB and other randomized benchmarks, utilize averages over circuits that contain random gates. Each time the circuit is run, a new gate is sampled from some ensemble, and the circuit outcomes are typically averaged together (so they are treated as though they came from the same circuit). For example, standard RB (see Sec. VIII B) uses sequences of random Clifford operations, whereas the randomized compiling [154,155] used in cycle benchmarking (Sec. VIII G) and Pauli noise learning techniques (Sec. IX C) inserts (and typically compiles in) random Pauli gates. At some point in the analysis of these techniques, one will encounter a superoperator A (see Sec. II C 2) that is averaged over all conjugations by elements of a group $\mathbb{G}$:

$$T_{\mathbb{G}}(A) = \int d\mu(g) gAg^{-1}, \tag{338}$$

where $d\mu$ is the Haar measure for the group $\mathbb{G}$. If $\mathbb{G}$ is a discrete group, then the Haar measure is just the counting measure, and the integral is often written as a sum:

$$T_{\mathbb{G}}(A) = \frac{1}{|\mathbb{G}|} \sum_{g\in\mathbb{G}} gAg^{-1}. \tag{339}$$

Equations (339) and (340) define the *twirl* of the superoperator A over the group $\mathbb{G}$. In both equations above, g is the superoperator (e.g., transfer matrix) representation of a group element g. So, even if we are twirling over a single-qubit unitary group, we will be using 4×4 transfer matrices, rather that the usual 2×2 unitary matrices. See Appendix C for a practical introduction to twirling and randomization.

We can understand group twirls by taking a brief diversion into representation theory. Recall that quantum operations act on density matrices (see Sec. II C). Transfer matrices are a representation of quantum operations that act on a vector space of *vectorized* density matrices (see Sec. II C 2). For a set of unitary superoperators that form a group $\mathbb{G}$, it turns out that we can divide this vector space into subspaces—irreducible representation spaces—in such a way that no element of $\mathbb{G}$ will mix distinct subspaces. This means that the entire set of superoperators in $\mathbb{G}$ can be simultaneously block diagonalized, with each block corresponding to an *irreducible representation*, or *irrep*. This decomposition into irreps is important, because Schur's lemma allows us to express the outcome of a twirl in terms of this decomposition. If the irreps are distinct (not related to each other by a similarity transform), then:

$$T_{\mathbb{G}}(A) = \sum_{\phi} \frac{\mathrm{Tr} A\mathbf{P}_{\phi}}{\mathrm{Tr}\mathbf{P}_{\phi}} \mathbf{P}_{\phi}. \tag{340}$$

where $\mathbf{P}_{\phi}$ is a projector onto the irreducible subspace of irrep ϕ.

The number and size of the irreps associated with the superoperator representation will depend on the group (representation) over which the twirl is being taken. The superoperator representation of the full unitary group $\mathsf{SU}(2^n)$ has just two irreps, a one-dimensional irrep that acts trivially on the trace of ρ, and a $(4^n - 1)$-dimensional irrep that mixes all other components. This means that the twirl of any superoperator A under the full unitary group will result in an n-qubit depolarizing channel—a diagonal Pauli transfer matrix (PTM; see Sec. II C 3) with a single unit eigenvalue and a real number $p \in [0, 1]$ repeated along

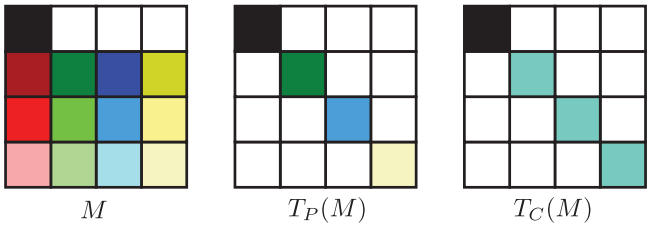


FIG. 22. Pauli and Clifford twirling. Twirling a PTM M with the Pauli group, $T_{\mathbb{P}}(M)$ simply eliminates the off-diagonal entries. Twirling with the larger Clifford group, $T_{\mathbb{C}}(M)$, also averages all but one (the top left) of the diagonal entries.

the rest of the diagonal. Importantly, this p is equal to the process polarization (see Sec. IV C 4) of A, i.e.,

$$p = f(A)\,. \tag{341}$$

Equivalently, A and $T_{\mathsf{SU}(2^n)}(A)$ have the same process (a.k.a. entanglement) fidelity to the identity [Eq. (235)]. So, an unknown error channel $\mathcal{E}$'s process fidelity can be learned by twirling it into a depolarizing channel and then learning that depolarizing channel's p, which is easy to do. This idea is foundational to RB.

The unitary group is an infinite group, and twirling over it, even approximately, can be experimentally challenging. So, often we consider twirls over smaller, discrete subgroups of the unitaries, such as the Clifford group, the Pauli group, or one of the dihedral groups. One consequence of twirling over a subgroup of the full unitary group is that the superoperator representation of a subgroup of $\mathsf{SU}(2^n)$ could decompose into significantly more irreps. The superoperator representation of the Clifford group actually breaks into the exact same irreps as the unitary group. Groups whose superoperators have the same irrep structure as the full unitary group are known as unitary 2-designs (see Appendix C 2), and are extremely important in QCVV. The superoperator representation of the n-qubit Pauli group, however, decomposes into 4^n one-dimensional irreps. Therefore, twirling a matrix M over the Pauli group will remove the off-diagonal entries of the matrix but leave the 4^n diagonal elements unchanged, as illustrated in Figs. 22 and 23. Twirling over the Clifford group will also project away the off-diagonal entries and will further replace all but one of the diagonal elements with their mean, as shown in Fig. 22.

B. Standard randomized benchmarking (RB)

There are many different RB protocols, and we cover many of them in this tutorial. But there is a *de facto* standard version of RB [284]—which we call *standard RB*—and we begin by explaining this protocol. This protocol is designed to benchmark any n-qubit gate set $\mathbb{G}_n$ (i.e., a set of n-qubit operations) that has the following two properties:

(1) $\mathbb{G}_n$ is a group (see Appendix B), and
(2) $\mathbb{G}_n$ is a unitary 2-design (see Appendix C 2).

The n-qubit Clifford group $\mathbb{C}_n$ has these properties, and it is almost always the gate set that is benchmarked using standard RB. We refer to standard RB with the Clifford group as *Clifford-group randomized benchmarking (CRB)*. Most standard RB experiments are one- or two-qubit CRB. For three or more qubits, more scalable RB protocols are typically used (see Sec. VIII C for further discussion). Note that the random circuits of standard RB (defined below) can be constructed for any gate set that is a group, but when that group is not a unitary 2-design, the data from those circuits will not have the simple form of standard RB, and a different data analysis procedure and/or different circuits are required (see Sec. VIII D).

Standard RB measures a mean error rate (i.e., average gate-set infidelity, or AGSI; see Sec. IV E 1) for the gates in $\mathbb{G}_n$, and it is designed so that the AGSI is not corrupted by SPAM errors (as long as the SPAM errors are not too large). It is given by the following protocol:

(1) Run $K \gg 1$ random motion-reversal circuits of L different circuit depths $m \geq 0$ and record each circuit's success frequency [285]. The circuit depths are typically linearly or logarithmically spaced, K is typically between 20 and 1000, $L \geq 3$ in order to fit an exponential function to the observed data (see below), and it is considered best practice to have all $m > 1$. Each of the K circuits at depth m is sampled and run as follows:

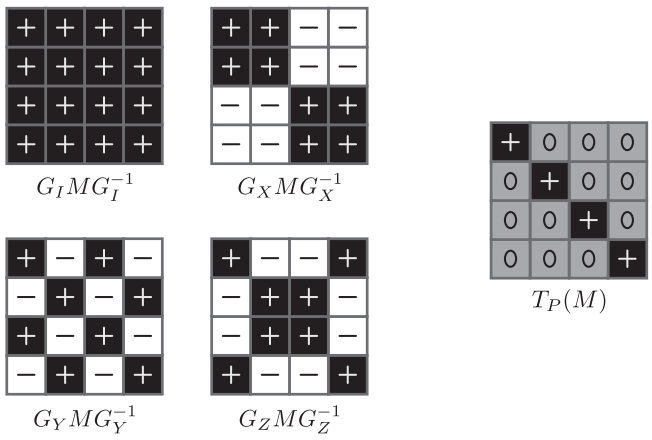


FIG. 23. Pauli Twirling. The PTMs for the Pauli operators are diagonal matrices of 1's and −1's. Conjugating a PTM M by a Pauli operator's PTM just changes the signs of some of the entries, as illustrated here for a single-qubit M and the four single-qubit Pauli operators. Averaging over all conjugations by Pauli operator PTMs, to get the twirled channel $T_P(M)$, preserves only the diagonal components.

(a) Uniformly and independently sample $m+1$ gates C from $\mathbb{G}_n$, $\{C_1, C_2, \ldots, C_{m+1}\}$, and construct a sequence $C_{m+1} \circ C_m \circ \cdots \circ C_2 \circ C_1$. This sampling can be done efficiently (in n) if $\mathbb{G}_n$ is the n-qubit Clifford group [286,287] (even though the size of $\mathbb{C}_n$ grows very quickly with n).

(b) Compute the inversion gate

$$C_{m+2} = P(C_{m+1} \circ C_m \circ \cdots \circ C_2 \circ C_1)^{-1}, \tag{342}$$

where P is an n-qubit Pauli operator. The original description of the standard RB protocol does not include P (i.e., it sets P to the identity). However, it is now considered best-practice to sample a uniformly random P [288–290]. Again, computing C_{m+2} is efficient if $\mathbb{G}_n$ is the n-qubit Clifford group [286].

(c) Construct a circuit $\mathcal{C}_m$ composed of the $m+1$ randomly sampled gates and the inversion gate:

$$\mathcal{C}_m = C_{m+2} \circ C_{m+1} \circ C_m \circ \cdots \circ C_2 \circ C_1 . \tag{343}$$

In the absence of errors, this circuit will always return the system to the original state, up to a final layer of Pauli gates determined by P. Thus, the ideal outcome is a particular bit string that is specified by P, which is the circuit's "success" outcome.

(d) Compile the circuit $\mathcal{C}_m$ into the native gates of the system being benchmarked, so that it can be measured experimentally. This compilation must simply replace each n-qubit Clifford in $\mathcal{C}_m$ with a sequence of those native gates that implements that particular unitary, i.e., "compilation barriers" must be placed between each layer in the circuit [291].

(e) Execute the compiled circuit $N \geq 1$ times and compute its success frequency:

$$p(\mathcal{C}_m) = N_{\text{success}}/N , \tag{344}$$

where N_{success} is the number of times the success outcome was observed. In experiments, typically N is between 100 and 1000. For a fixed value of $K \times N$ (which is the total number of circuit executions in the RB experiment), $N = 1$ is statistically optimal [292], i.e., it results in the lowest uncertainties on the AGSI estimated by RB. However, due to the time required to compile circuits and upload waveforms in most experimental setups, sufficiently low uncertainty estimates of the AGSI can typically be achieved most quickly by setting $N \gg 1$. Each circuit execution is the following procedure:

(1) Prepare each of the n qubits in the $|0\rangle\langle 0|$ state.
(2) Apply the circuit $\mathcal{C}_m$.
(3) Measure all n qubits in the computational basis, and check whether the "success" bit string was observed.

(2) Compute the average success probability $\bar{p}(m)$ for each depth m,

$$\bar{p}(m) = \frac{1}{K} \sum_{\mathcal{C}_m} p(\mathcal{C}_m) . \tag{345}$$

Then, fit this data to an exponential decay function:

$$\bar{p}(m) = Af^m + B , \tag{346}$$

where A, f, and B are fit parameters. If the success bit string has been randomized (i.e., P is uniformly random), fix $B = 1/2^n$ (which provides a higher-precision estimate of f for the same amount of data [288–290]). A is typically called the "SPAM parameter," because it includes information about the size of the SPAM errors [293]. RB data is typically analyzed with simple curve fitting routines (e.g., weighted least squares), although there are a variety of alternative fitting approaches. Note that meaningful results will only be obtained if $\bar{p}(m_{\text{min}})$ is significantly greater than $\bar{p}(m_{\text{max}})$, where m_{min} and m_{max} are the minimum and maximum depths used, which requires that the SPAM errors are not catastrophically large.

(3) RB theory shows that under certain circumstances (see below) the fit f is an estimate of the mean process polarization [Eq. (245)] of the gates in $\mathbb{G}_n$, but it is more common to report (in)fidelities than polarization. An estimate of the *mean* of the gates' infidelities is given by the average gate infidelity [Eq. (223)],

$$r = \frac{d-1}{d}(1-f) , \tag{347}$$

or the process (i.e., entanglement) infidelity [Eq. (237)],

$$e_F = \frac{d^2-1}{d^2}(1-f) , \tag{348}$$

where $d = 2^n$ is the dimension of the Hilbert space for n qubits (see Table II for a summary of the linear relationships between these different quantities). r or e_F is an estimate of the mean of the infidelities

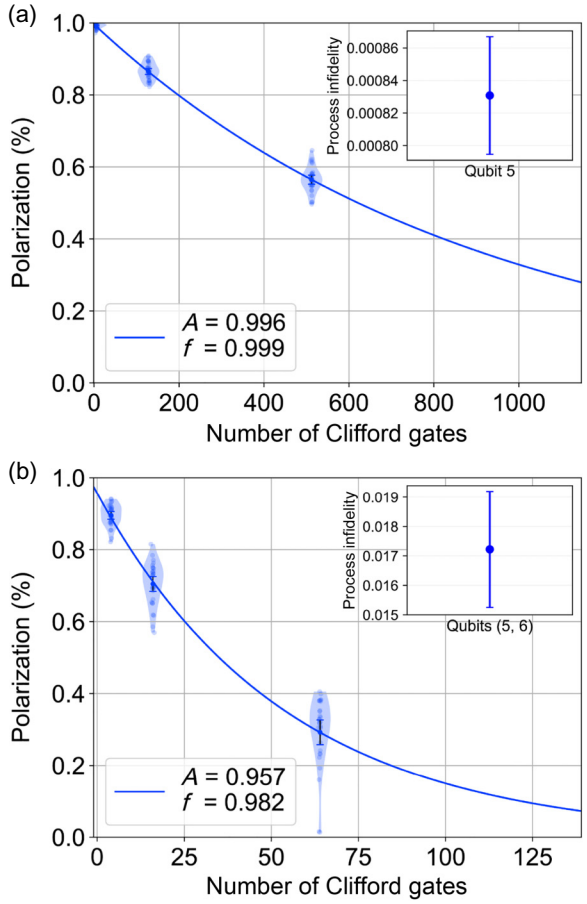


FIG. 24. One- and two-qubit Clifford-group RB. Exponential decays for CRB performed on (a) a single qubit (labeled 5) and (b) two qubits (labeled 5 and 6) on a superconducting quantum processor. The results are plotted in terms of the polarization [Eq. (350)], a rescaling of the success probability. For both experiments, $K = 30$ random Clifford circuits were generated for each circuit depth m at $L = 3$ different circuit depths. The SPAM parameters A and exponential fit parameters f are listed in the legends. At each circuit depth, circular data points depict the results of individual circuits and violin plots depict the distribution of results. Insets: the process infidelity for the (a) single-qubit CRB results [$e_F = 8.3(2) \times 10^{-4}$] and (b) two-qubit CRB results [$e_F = 1.7(1) \times 10^{-2}$]. The exponential decay curve for two-qubit CRB decays much faster than for single-qubit CRB, demonstrating that the error per Clifford is larger for two-qubit Cliffords than single-qubit Cliffords, as shown by their relative process infidelities.

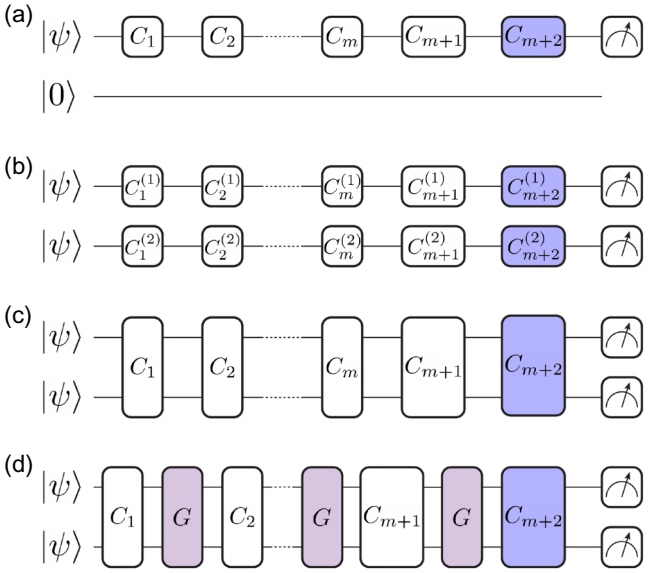


FIG. 25. Standard, simultaneous, and interleaved RB. (a) Circuit structure for standard single-qubit RB. Gates are only applied to the benchmarked qubit; all other qubits are assumed to remain in their ground states. (b) Circuit structure for simultaneous RB. Here, gates are applied to two or more qubits simultaneously to benchmark the performance of simultaneous gate operations, where $C_m^{(i)}$ denotes the mth gate applied to the ith qubit. (c) Circuit structure for standard two-qubit RB. (d) Circuit structure for interleaved RB. Here, G (purple) is the interleaved gate whose infidelity can be estimated from analyzing (c),(d) together (see Sec. VIII F). For (a)–(d), C_{m+2} (blue) is the inversion gate for the entire sequence.

of the gates in $\mathbb{G}_n$, so in the case of CRB this is often called the *error per Clifford (EPC)*. When comparing RB error rates, it is important to check whether the convention in Eq. (348) or (349) is being used, as the process infidelity is stable under tensor products of parallel gates, while the average gate infidelity is not (see the discussion in Sec. IV C 3 b).

Examples of results for one-qubit and two-qubit CRB experiments are shown in Fig. 24, and the forms of the circuits used in one- and two-qubit CRB are shown in Figs. 25(a) and 25(c), respectively. In these figures, we plot *polarization* rather than success probability. The polarization S_{pol} is simply a rescaling of success probability (p), given by

$$S_{\text{pol}} = \frac{p - 1/2^n}{1 - 1/2^n}. \tag{349}$$

Polarization is sometimes more convenient because $S_{\text{pol}} = 0$ when all n qubits are completely depolarized [294]. The measured one-qubit and two-qubit EPCs are $e_F = 8.3(2) \times 10^{-4}$ and $e_F = 1.7(1) \times 10^{-2}$, respectively. In most systems, it is expected that the EPC for two-qubit CRB will be higher than single-qubit CRB, since two-qubit CRB requires two-qubit entangling gates, which are typically noisier than single-qubit gates.

To run n-qubit CRB experiments, each n-qubit Clifford operation must be decomposed into the system's native gates (step 1d above). For example, in a widely used compilation [295] of the 24 single-qubit Clifford gates ($\mathbb{C}_1$) into rotations around X and Y, the average number of

single-qubit native gates per single-qubit Clifford gate is 1.875. CRB estimates the average error rate of the composite n-qubit Clifford gates (the EPC), *not* the average error rate of the fundamental gates from which those gates are composed, and the EPC depends on the compilation used. However, it is common practice to rescale single-qubit CRB's EPC ($r_{\mathbb{C}_1}$) to a native gate error rate. For example, for the compilation of Ref. [295], the EPC is typically related to the error per native single-qubit gate (r_{SQ}) with the simple heuristic:

$$r_{\mathbb{C}_1} = 1.875 r_{\mathrm{SQ}} \,. \tag{350}$$

An alternate heuristic for estimating the error per native single-qubit gate is to use the common compilation strategy of decomposing all U_3 gates (i.e., arbitrary single-qubit rotations parametrized by three Euler angles) into a sequence consisting of three parametrized *virtual* Z gates and two *physical* native $X_{\pi/2}$ gates [296]:

$$U_3(\phi,\theta,\lambda) = Z_{\phi-\frac{\pi}{2}} X_{\frac{\pi}{2}} Z_{\pi-\theta} X_{\frac{\pi}{2}} Z_{\lambda-\frac{\pi}{2}} \,. \tag{351}$$

Now, there are always two real gates (i.e., physical pulses) per single-qubit Clifford [297], and thus the EPC is twice the error per native gate. One advantage of this approach is that it is straightforward to generalize to higher dimensions [298], and has been used to estimate native gate fidelities in single-qutrit ($d = 3$) and single-ququart ($d = 4$) CRB experiments [272], where 6 and 12 native gates are needed per single-qutrit and single-ququart Clifford gate, respectively; see Appendix D 3 for an overview of randomized benchmarks for qudits.

Similarly, for two-qubit CRB and the widely used compilation of Ref. [295], a two-qubit Clifford gate contains 1.5 CNOT or CZ gates and 8.25 single-qubit gates, on average. For this compilation, two-qubit CRB's EPC ($r_{\mathbb{C}_2}$) is then often related to r_{SQ} and the two-qubit gate error rate (CZ) using the simple heuristic

$$r_{\mathbb{C}_2} = \frac{3}{2} r_{\mathrm{CZ}} + \frac{33}{4} r_{\mathrm{SQ}} \,. \tag{352}$$

These rescalings of the EPC are only heuristics though—they are known to *not* reliably estimate the native gate infidelities, in general. Importantly, the estimated error per native gate will typically change if different compilations are used. This is because CRB circuits prevent systematic addition or cancellation of coherent errors between different n-qubit Clifford gates, but not within the gate sequences used to create each n-qubit Clifford gate.

We now explain how to interpret RB results, and why RB works, by concisely summarizing the practical implications of the theory of standard RB. Standard RB works because the random gates twirl the errors in the gates, and because each gate is sampled from a unitary 2-design (such as the Clifford group) this twirl maps the gates' (potentially complicated) error maps into depolarization channels (as outlined in Sec. VIII A). Turning this into a precise theory for RB is simple in the "gate-independent noise" idealization, where every gate in $\mathbb{G}_n$ is subject to the same CPTP error map $\mathcal{E}$. In this case, it is possible to show that standard RB's average success probability satisfies

$$\bar{p}(m) = A f(\mathcal{E})^{m+1} + B \,. \tag{353}$$

Here, $f(\mathcal{E})$ is $\mathcal{E}$'s process polarization, and A and B absorb all SPAM error (and also have contributions from gate error). Straightforward derivations of this equation can be found throughout the literature on RB theory (e.g., see the "zeroth-order model" in Ref. [299]). Therefore, for gate-independent noise, r [Eq. (348)] or e_F [Eq. (349)] is a rigorous estimate of the average gate infidelity or process infidelity of each gate's error map $\mathcal{E}$, respectively.

Understanding RB outside of the unrealistic setting of gate-independent noise is more complex. The modern theory of RB [196,198,300,301] addresses the more realistic setting in which each gate has its own distinct error map. We will not delve into this theory here, but we highlight its main practical implications:

(a) Standard RB's average success probability will decay exponentially as long as the gates experience only moderately small Markovian errors [196,198, 300,301] and the minimal circuit depth (m) used is sufficiently large. The theory shows that, under typical circumstances, $m \geq 1$ (corresponding to three Clifford gates in our depth convention) is sufficient. Therefore, standard RB data that is inconsistent with an exponential decay implies the presence of non-Markovian errors. For example, $1/f$ noise is well-known to cause nonexponential RB decays [288].

(b) The simplest interpretation of standard RB's f parameter is that it is equal to the mean of the gate's process polarizations, and therefore r (or e_F) is equal to the mean of the gates' infidelities, i.e.,

$$\epsilon_{\mathrm{uni}} = \frac{1}{|\mathbb{G}_n|} \sum_{G \in \mathbb{G}_n} \epsilon(G) \,, \tag{354}$$

where $\epsilon(G)$ is the average gate or process infidelity of G. Unfortunately, although this interpretation contains the essence of what r measures [302], it is subtly incorrect (in part because ϵ_{uni} is ill defined, due to gauge ambiguities [196]). A mathematically precise understanding of r's relationship to gate infidelity is not important for using RB. But it *is* practically relevant when checking whether concurrent RB and tomography experiments have consistent results. Correctly predicting r from measured

transfer or process matrices requires either (a) using modern RB theory's predictions for how to compute r from transfer or process matrices [196,198,300,301], or (b) simply simulating RB experiments using those transfer or process matrices.

C. Native gate RB

Native gate RB is a family of methods that can directly benchmark a system's native n-qubit gates (which are often referred to as "layers" or "cycles," but here we will follow RB convention and call them "gates"). The main native gate RB techniques are as follows:

(a) direct RB (Sec. VIII C 1),
(b) binary RB (Sec. VIII C 2),
(c) mirror RB (Sec. VIII C 3), and
(d) linear cross-entropy benchmarking (Sec. VIII C 4).

Native gate RB protocols address two practical limitations of standard CRB. Firstly, CRB is infeasible beyond a few qubits even with state-of-the-art gate-error rates. This is because CRB runs circuits containing uniformly random elements of the n-qubit Clifford group, and the size of the circuits needed to create these n-qubit Clifford gates grows very rapidly with n for typical native gate sets. In particular, a typical n-qubit Clifford gate requires $\mathcal{O}(n^2/\log n)$ two-qubit gates [303–307]. The average success probability of even the shortest CRB circuits therefore quickly drops off to almost zero as n increases [308], as demonstrated in Fig. 27(a). This makes it impossible to estimate the EPC without impractical amounts of data when $n \gg 1$ (and the EPC rapidly converges to 1 as n increases). Secondly, CRB measures the EPC, but most users of RB actually want to know the error per native gate. Although rescaling the EPC to estimate the error per native gate is common practice (see the discussion in the previous subsection), it has little theoretical justification [309]. Furthermore, beyond the one- and two-qubit setting, it is not even typically clear what would constitute a sensible and useful rescaling of the EPC.

Native gate RB protocols benchmark some user-specified set of n-qubit gates $\mathbb{G}_n$. In all the protocols described herein, this gate set is required to *generate* a group that is a unitary 2-design, such as the Clifford group. A one-qubit example of such a gate set is

$$\mathbb{G}_1 = \left\{X_{\frac{\pi}{2}}, Y_{\frac{\pi}{2}}\right\}. \tag{355}$$

In experimental uses of native gate RB methods to date, $\mathbb{G}_n$ has typically been chosen to be parallel applications of either (a) a system's native gates, or (b) one- and two-qubit gates that can easily be constructed from the native gates (e.g., all possible layers consisting of parallel applications of CNOT and single-qubit Clifford gates). Other choices for $\mathbb{G}_n$ are possible, however.

Native gate RB protocols estimate an average error rate (r_Ω) for the gates in $\mathbb{G}_n$ that is weighted by a user-specified probability distribution Ω over $\mathbb{G}_n$. This error rate is, in essence, the Ω-weighted average infidelity of the gates, i.e.,

$$\epsilon_\Omega = \sum_{G\in\mathbb{G}_n} \Omega(G) e_F(G), \tag{356}$$

where $e_F(G)$ is the process infidelity of G [308,310–312] (although, as with CRB, there are some subtleties relating r_Ω to ϵ_Ω because ϵ_Ω is not gauge-invariant [308,311]). The distribution Ω can be chosen to measure the weighted error rate of most interest, and can even be varied to learn about which gates have higher error rates [310,311]. For the gate set example in Eq. (356), an given of such a distribution is

$$\Omega(X_{\frac{\pi}{2}}) = 3/4, \quad \Omega(Y_{\frac{\pi}{2}}) = 1/4. \tag{357}$$

All native gate RB protocols follow a similar procedure to standard RB. Stated informally, they all have the following structure:

(1) Run random circuits of various depths m. The exact structure of the random circuits varies between different methods (see Fig. 26), but in all cases the circuits consist of
   (a) m random layers sampled from a user-specified distribution Ω, called Ω-*distributed random circuits*, surrounded by
   (b) some additional, method-specific state-preparation and measurement layers (or subcircuits).
(2) Estimate a success metric for each circuit, the details of which depend on the protocol [313].
(3) Estimate the average of this success metric at each depth, fit that average to the exponential decay function of Eq. (347), and from the fitted value for f compute an error rate r_Ω using Eq. (348).

Native gate RB protocols work because random circuits randomize and spread errors (i.e., via "scrambling") [308]—which must happen because a random sequence of elements from $\mathbb{G}_n$ converges to a random element generated from the group $\mathbb{G}_n$, which is a unitary 2-design. The implication of this is that the process fidelity of Ω-distributed random circuits ($F_{\Omega,d}$) will decay exponentially with circuit depth at a rate given by ϵ_Ω under broad conditions [302,308,311]. Each native gate RB protocol differs in (i) the state preparation and measurement structures used in its circuits, and (ii) its choice of success metric. These differences correspond to different ways to measure $F_{\Omega,d}$, each of which has its own strengths and weaknesses.

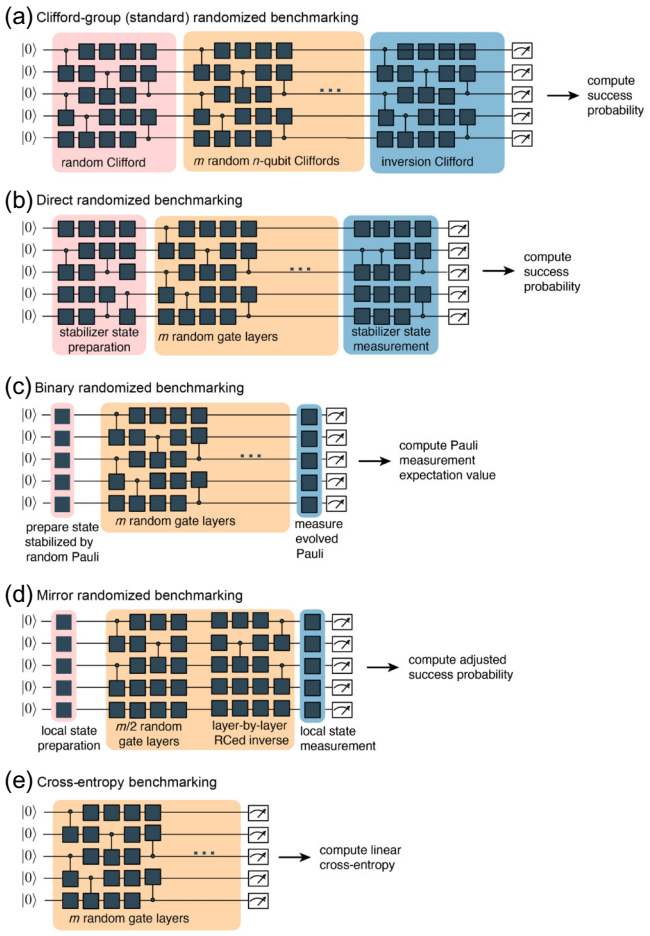


FIG. 26. Clifford-group and native gate RB methods. The structure of the random circuits and the success metrics used in (a) Clifford-group RB, (b) direct RB, (c) binary RB, (d) mirror RB, and (e) cross-entropy benchmarking. The red denotes state-preparation layers (or subcircuits), orange denotes the benchmarking sequence, and blue denotes the measurement basis rotations. CRB circuits consist of $m+1$ uniformly random n-qubit Clifford gates followed by the unique Clifford gate that inverts those $m+1$ gates, and so CRB measures the mean error rate of these n-qubit Clifford gates, often called the "error per Clifford" (EPC). The native gate RB protocols (b)–(e) are all based on circuits containing m layers of gates sampled from some distribution Ω over a layer set $\mathbb{G}_n$, and most of them surround that "Ω-distributed random circuit" with additional circuits that implement different state preparations and measurements. The layer set used in these protocols is typically closely related to the set of all native layers for a system. These techniques all measure an error rate r_Ω that quantifies the Ω-weighted average error rate of these layers. Each of the native gate RB protocols has its own strengths, limitations, and regimes of applicability, which are discussed in the main text.

1. Direct RB

Direct randomized benchmarking (DRB) [308,310] can benchmark any gate set that generates a group that is a unitary 2-design. It has been primarily used to benchmark gate sets that generate the n-qubit Clifford group [310,314,315], and so we focus on that case. The circuits used in DRB (i) begin with a random subcircuit that creates a uniformly random stabilizer state, (ii) have a depth m Ω-distributed circuit at their center, and (iii) end with a subcircuit that maps the state that is (ideally) produced by the circuit so far to a random computational basis state. The structure of DRB circuits is shown in Fig. 26(b).

Each DRB circuit always outputs a particular bit string b when run without error, and the probability that this bit string is observed is DRB's success metric. The initial and final subcircuits within a DRB circuit implement a (state) 2-design twirl on the error in the Ω-distributed circuit. This guarantees that the mean success probability of DRB circuits decays exponentially and DRB's error rate (r_Ω) approximately equals the weighted-average error rate ϵ_Ω of the benchmarked gates [308,310].

Figure 27(b) demonstrates DRB. This figure shows the average polarization decay obtained when running n-qubit DRB experiments on an `IBM Q` system, for $n = 1$ to 6. The key differences between DRB and CRB are illustrated by comparing Fig. 27(b) to the results of n-qubit CRB experiments run at the same time on the same system, shown in Fig. 27(a). First, the CRB polarization decays more quickly with depth than the DRB polarization does, i.e., they are measuring different error rates. Importantly, CRB measures the EPC, whereas DRB measures the error per layer of native gates. Second, the average polarization of the shallowest DRB circuits ($m = 0$) is typically larger than that of the shallowest CRB circuits (also $m = 0$). This is because generating a uniformly random stabilizer state requires only about 1/3 as many two-qubit gates as generating a uniformly random Clifford gate [306]. This means that DRB is feasible on more qubits than CRB. However, DRB is still not truly scalable, because DRB's state preparation and measurement subroutines require $\mathcal{O}(n^2/\log n)$ two-qubit gates [303–307]. These large subroutines mean that the polarization of $m = 0$ DRB circuits still drops rapidly with increasing n [see Fig. 26(b)], even though it does not drop as quickly as CRB.

2. Binary RB

Binary randomized benchmarking (BiRB) [312] is a native gate RB protocol that is designed to benchmark any gate set that generates the n-qubit Clifford group. BiRB's circuit structure is shown in Fig. 26(c). Unlike most RB protocols, BiRB's circuits do not include an inversion gate or subcircuit at their end—i.e., they are not motion reversal circuits. BiRB circuits are therefore not definite outcome circuits, since they do not always return a particular "success" bit string if run without error. Instead, BiRB circuits consist of an Ω-distribution random circuit with a layer of single-qubit gates at its start and at its end, and 50% of all possible n-bit strings are designated as "success" bit strings and the other 50% as "fail" bit strings.

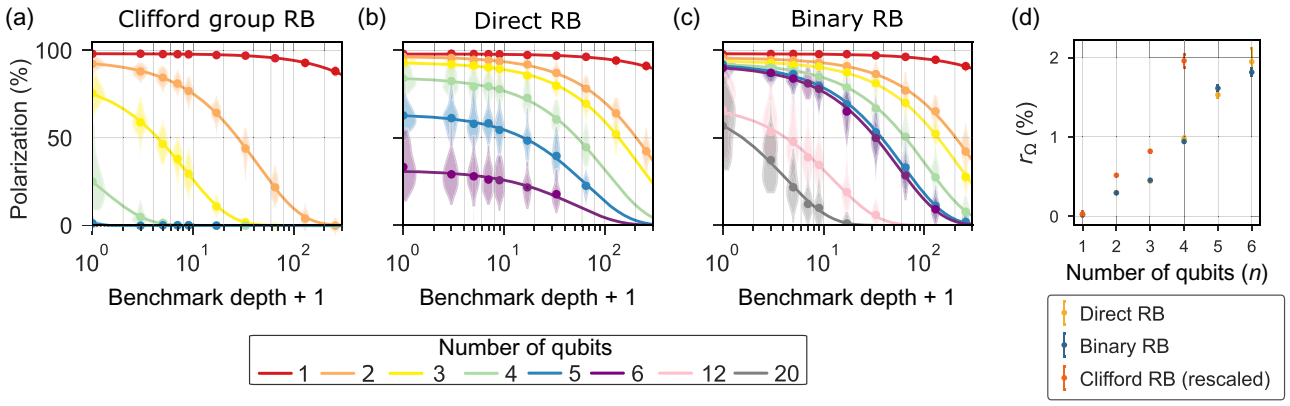


FIG. 27. Comparing Clifford-group RB, direct RB, and binary RB. The results of running (a) Clifford-group RB, (b) direct RB, and (c) binary RB on `ibm-hanoi`. For all protocols, the $m = 0$ polarization drops as the number of benchmarked qubits increases, but this effect is smaller for protocols with shorter state preparation and measurement subroutines. (d) The error rates extracted from each dataset. DRB and BiRB both measure the error rate of layers sampled from a distribution Ω, and their results are similar. In contrast, CRB measures the error per n-qubit Clifford (EPC), which is often rescaled to estimate the error rate of native gates, but is not guaranteed do so accurately. (Figure adapted with permission from Ref. [312].)

The initial layer of single-qubit gates in BiRB circuits creates a tensor product eigenstate of a uniformly random n-qubit Pauli operator P. This simulates sending a uniformly random Pauli operator P into an Ω-distributed circuit—a technique that enables scalable fidelity estimation, and which is also used in direct fidelity estimation (Sec. IX B), cycle benchmarking (Sec. VIII G), and Pauli noise learning (Sec. IX C). In the absence of errors, the Ω-distributed circuit transforms P into another Pauli operator P'. The final layer of gates simply transforms P' into a Z-type Pauli operator P'' (a tensor product of Z and I operators), enabling the measurement of whether P "survived" the circuit (i.e., was correctly transformed by the circuit) using only a computational basis measurement. If the readout bit string is a $+1$ eigenstate of P'' we declare "success," and otherwise we declare "fail." We then (i) compute the success metric

$$p = \nu_{\text{success}} - \nu_{\text{fail}}\,, \tag{358}$$

where ν_{success} and ν_{fail} are the frequencies at which success and fail bit strings are observed, respectively; (ii) fit the mean of p versus depth to the standard exponential decay function of Eq. (347) with $B = 0$; and (iii) estimate r_Ω using Eq. (348).

An example of BiRB data is shown in Fig. 27(c). BiRB is more scalable than DRB (and CRB) and it is also arguably simpler to implement. This is because BiRB's circuits do not start and end with large subcircuits [see Fig. 26(c)]. The better scaling of BiRB can be seen by comparing Fig. 27(c) with (a) and (b). This shows that the polarization of the shallowest BiRB circuits decreases more slower than that of both DRB and CRB circuits as a function of the number of qubits. Note, however, that there is still a gradual decrease in the polarization of these shallowest circuits due to the increasing SPAM error with n.

3. Mirror RB

Mirror randomized benchmarking (MRB) [206,311,316,317] is a native gate RB protocol that is scalable because it uses random "mirror circuits" [294] (see Secs. IX C 2, X A, and XI A 2 for more on mirror circuits). MRB can efficiently benchmark both Clifford and universal gate sets. The structure of MRB circuits is shown in Fig. 26(d). A MRB circuit of benchmark depth m consists of (i) a layer of single-qubit gates each sampled independently from a 2-design, (ii) $m/2$ gates sampled from Ω, and (iii) a depth-$(1 + m/2)$ circuit consisting of each layer in the circuit so far, but in the reverse order, and each replaced with its inverse (i.e., a layer-by-layer inversion circuit). Randomized compiling is applied to the entire circuit. Randomized compiling ensures that errors do not coherently add or cancel between a layer and its inverse in the second half of the circuit [206,294,311]. Note that unlike in DRB, BiRB, and linear cross-entropy benchmarking (Sec. VIII C 4), the m layers are *not* all sampled independently from Ω. Instead, $m/2$ layers are sampled independently from Ω, and then the next $m/2$ layers are their inverses. MRB's use of a layer-by-layer inverse removes the large state-preparation and measurement subroutines used in DRB (and CRB) circuits.

MRB's state preparation and measurement layer of single-qubit gates are based on the insight that the infidelity of an error channel can be efficiently estimated using single-qubit 2-design twirling. However, this requires a more complex success metric than the frequency of observing the "success" bit string, used in DRB and CRB. In

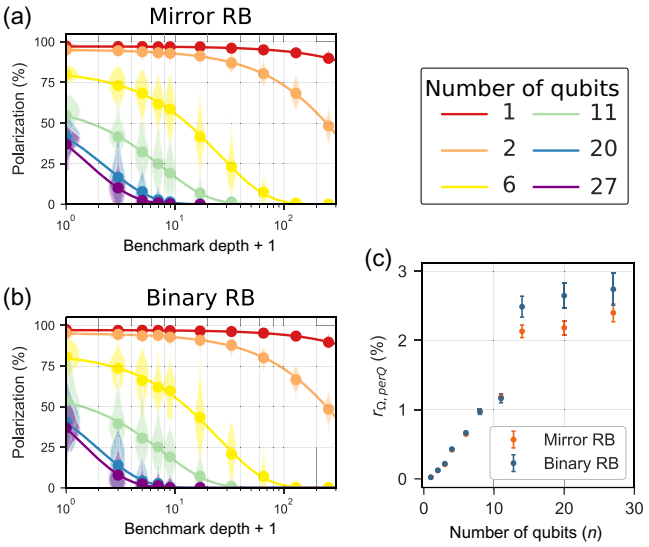


FIG. 28. Demonstrating mirror RB and binary RB. Results from running (a) mirror RB and (b) binary RB on `ibm-kolkata` to benchmark a Clifford gate set. (c) The RB error rate per qubit (approximately r_Ω/n) extracted from each dataset. The error rate per qubit increases with n, which indicates the presence of crosstalk. MRB and BiRB are both designed to measure the error rate of layers sampled from a distribution Ω, but MRB is known to typically slightly underestimate this error rate, which is consistent with the observations here. However, MRB can efficiently benchmark a universal gate set, whereas no other RB protocol can. (Figure adapted with permission from Ref. [312].)

MRB, the success metric is based on the Hamming distance of the observed bit string from the success bit string. In particular, MRB's success metric—called the *adjusted success probability*—is

$$p = \frac{4^n}{4^n-1}\left[\sum_{k=0}^{n}\left(-\frac{1}{2}\right)^k h_k\right] - \frac{1}{4^n-1}, \tag{359}$$

where h_k is the frequency that the circuit outputs a bit string with Hamming distance k from its target bit string. The theories in Refs. [160,206,318] show that Eq. (360) is a reliable estimator of fidelity when using a local 2-design twirl.

Figure 28 demonstrates the use of MRB to benchmark a set of layers that generate the Clifford group, and compares MRB to BiRB of the same layer set. The correlations in MRB circuits enable creating motion reversal circuits without large "overhead" subroutines, as in DRB circuits, but they also have an unwanted side effect. MRB theory [206,311] shows that if the error rates of a Ω-distributed layer and its inverse are uncorrelated, then MRB accurately estimates ϵ_Ω, but that if these error rates are correlated then MRB *slightly* underestimates ϵ_Ω. In real systems, these error rates typically are correlated, resulting in MRB slightly underestimating ϵ_Ω [206,311]. We observed this effect in Fig. 28, with BiRB's error rates slightly larger the MRB's error rates. BiRB is, therefore, expected to estimate ϵ_Ω marginally more accurately than MRB, and BiRB is just as scalable as MRB. However, MRB can efficiently benchmark universal gate sets (e.g., see the experiments in Ref. [311]), where as BiRB (and DRB) cannot.

4. Cross-entropy benchmarking

Cross-entropy benchmarking (XEB) is a collection of related protocols that run random circuits and quantify how well they performed by estimating the cross-entropy between the actual ($\mathbf{q}$) and ideal ($\mathbf{p}$) outcome distributions [2,136,319–322]. In practice, these techniques typically use the *linear* cross-entropy [see also Eq. (190)]:

$$H_{\mathrm{lin}}(\mathbf{p},\mathbf{q}) \equiv 2^n \sum_x p_x q_x - 1, \tag{360}$$

where the sum is over all n-bit strings.

In the context of QCVV, the most important XEB methods are a family of protocols for measuring the average error rate (ϵ_Ω) of n-qubit circuit layers and gates—i.e., they measure the same quantity as other native gate RB protocols discussed throughout Sec. VIII C—and this is the type of protocol we detail below. But, first, we briefly discuss another meaning for "XEB"—the protocol used for demonstrating "quantum supremacy" [2,136]. That XEB procedure is as follows: (i) run the n-qubit scrambling circuits described in the "quantum supremacy" literature [2,136], (ii) run experiments to estimate $H_{\mathrm{lin}}(\mathbf{q},\mathbf{p})$, where $\mathbf{q}$ is the actual and $\mathbf{p}$ the ideal outcome distributions for each sampled circuit (note that this estimation is challenging when $\mathbf{p}$ is infeasible to compute with classical simulations of the circuit), and (iii) use the value of this cross-entropy to quantify a quantum computer's performance. For sufficiently deep circuits on sufficiently many qubits, values of $H_{\mathrm{lin}}(\mathbf{p},\mathbf{q})$ above some threshold δ are believed to be impossible to achieve in a reasonable amount of time using any existing classical computer [2–4]. Obtaining such values for $H_{\mathrm{lin}}(\mathbf{p},\mathbf{q})$ is sometimes referred to as demonstrating "quantum supremacy," and this has now been achieved in multiple experiments [2–4].

We now turn to the XEB protocols that are designed to estimate the average infidelity of random n-qubit circuit layers (ϵ_Ω). These XEB protocols follow the same structure as all other native gate RB protocols (discussed throughout this subsection), using (i) plain Ω-distributed random circuits [see Fig. 26(e)] as its circuit family and (ii) a success metric related to the linear cross-entropy. The success metric is typically

$$F_{\mathrm{XEB}} = \frac{\hat{H}_{\mathrm{lin}}(\mathbf{p},\mathbf{q})}{H_{\mathrm{lin}}(\mathbf{p},\mathbf{p})}, \tag{361}$$

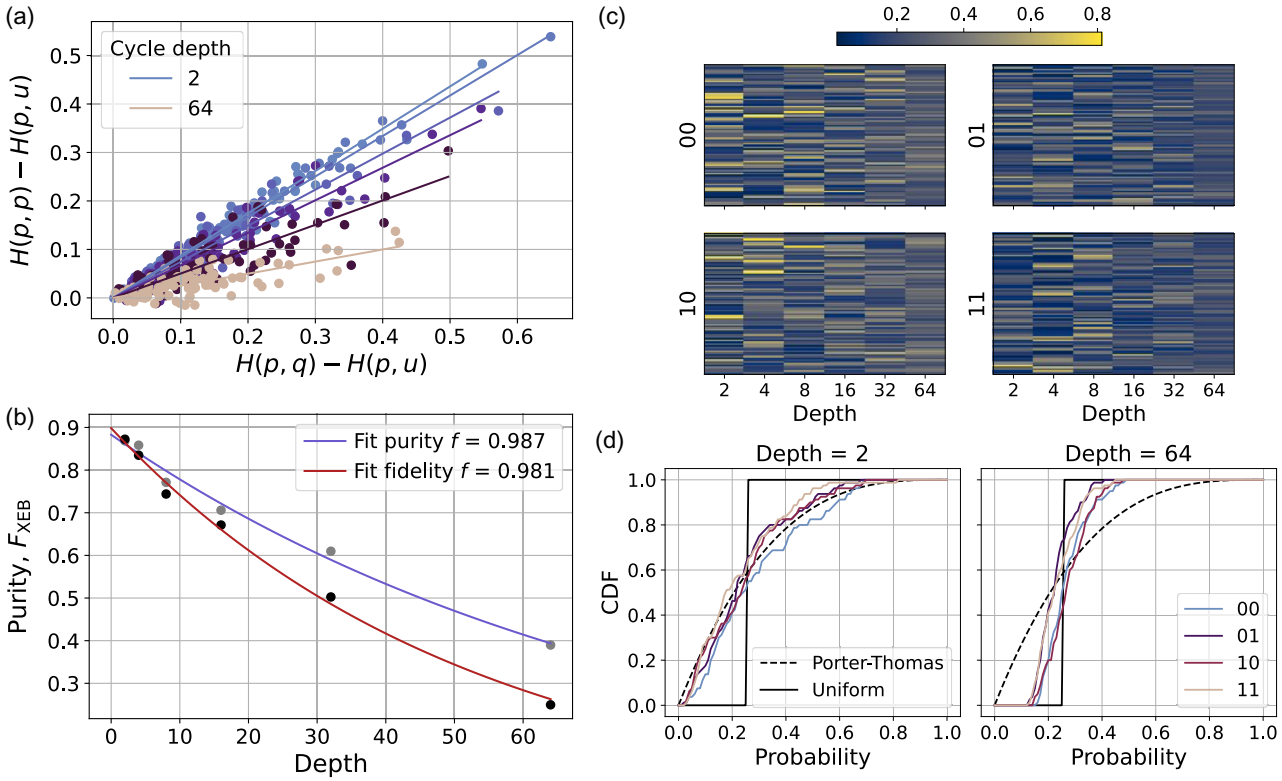


FIG. 29. Cross-entropy benchmarking of a two-qubit CZ gate. (a) Linear fits yielding $F_{\mathrm{XEB}}(m)$ at each depth m. (b) Exponential decay fits (solid lines) of measured (dots) $F_{\mathrm{XEB}}(m)$ (black) and speckle purity (gray) as a function of depth m. The dressed CZ gate has a process infidelity of $e_F = 1.78\%$, of which 68% can be attributed to stochastic errors based on the decay of the speckle purity. (c) The speckle pattern is plotted for each bit string across $N = 30$ random circuits (y axis) and at each depth m (x axis). We see that the speckled pattern at low depths (characteristic of the Porter-Thomas distribution) is smeared out at longer depths (as the uniform distribution is approached). (d) Cumulative distribution function (CDF) of distributions of bit string probabilities (colored lines) at depths $m = 2$ and $m = 64$. At low depth, $m = 2$, the probabilities of the various bit strings closely follow the Porter-Thomas distribution (dashed black line). At larger depths these distributions begin to converge to the uniform distribution (solid black line) where the probability of the given bit string is close to $1/d = 0.25$ for all of the $N = 30$ random circuits.

where $\hat{H}_{\mathrm{lin}}(\mathbf{p},\mathbf{q})$ is an estimate of $H_{\mathrm{lin}}(\mathbf{p},\mathbf{q})$. Typically, the estimate is computed using

$$\hat{H}_{\mathrm{lin}}(\mathbf{p},\mathbf{q}) = \frac{2^n}{N}\sum_{x\in\mathbb{X}} p_x - 1\,, \tag{362}$$

where $\mathbb{X}$ is the set of bit strings observed when running the circuit N times. Calculating Eq. (363) requires classically computing p_x for each $x \in \mathbb{X}$, which is generally exponentially expensive in the number of qubits. In the most well-known XEB experiments [2], the mean of F_{XEB} over multiple random circuits of depth m—given $\hat{H}_{\mathrm{lin}}(\mathbf{p},\mathbf{q})$ and $H_{\mathrm{lin}}(\mathbf{p},\mathbf{p})$ for each circuit—is estimated by plotting $\hat{H}_{\mathrm{lin}}(\mathbf{p},\mathbf{q}) - \hat{H}_{\mathrm{lin}}(\mathbf{p},\mathbf{u})$ versus $H_{\mathrm{lin}}(\mathbf{p},\mathbf{p}) - \hat{H}_{\mathrm{lin}}(\mathbf{p},\mathbf{u})$ for every circuit of the same depth, where $\mathbf{u}$ is the uniform distribution in d dimensions, and fitting that data to a line [as shown in the example of Fig. 29(a)]. Note, however, that alternative success metrics or data analysis techniques can be used in XEB; for example, the success metric defined by Eq. (361) can be used instead of Eq. (362), or the techniques of *filtered RB* [301,321] can be applied to data from XEB circuits.

XEB's success metric is arguably less intuitive than the success probability used in most RB protocols, and so we now explain why F_{XEB} enables estimating the average error rate of the benchmarked layers. Consider a depth-m XEB circuit $\mathcal{C}$ and the observable $O_{\mathcal{C}} = \sum_x p_x |x\rangle\langle x|$, where $\mathbf{p} = \{p_x\}$ is $\mathcal{C}$'s ideal outcome distribution. Now, assume that $\mathcal{C}$'s imperfect implementation can be modelled by an n-qubit depolarizing error channel after each layer with process polarization f, i.e., the state output by $\mathcal{C}$ is

$$\rho_{\mathcal{C}} = f^m |\psi_{\mathcal{C}}\rangle\langle\psi_{\mathcal{C}}| + (1-f^m)\frac{\mathbb{I}}{2^n}\,, \tag{363}$$

where m is the circuit's depth, and $|\psi_{\mathcal{C}}\rangle\langle\psi_{\mathcal{C}}|$ is the pure state that $\mathcal{C}$ would ideally create. Then,

$$\mathrm{Tr}[O_{\mathcal{C}}\rho_{\mathcal{C}}] = f^m \mathrm{Tr}[O_{\mathcal{C}}\psi_{\mathcal{C}}] + (1-f^m)\frac{1}{2^n}\mathrm{Tr}[O_{\mathcal{C}}]\,. \tag{364}$$

By substituting in O_C, we find that

$$\sum_x p_x q_x = f^m \left(\sum_x p_x^2 - \frac{1}{2^n} \right) + \frac{1}{2^n}, \tag{365}$$

where $q_x = \langle x | \rho_C | x \rangle$ is the actual probability of observing x. Rearranging, and substituting in the definition of the linear cross-entropy, we obtain

$$f^m = \frac{\hat{H}_{\text{lin}}(\mathbf{p}, \mathbf{q})}{H_{\text{lin}}(\mathbf{p}, \mathbf{p})}. \tag{366}$$

So, by estimating the RHS of this equation (for randomly sampled circuits) versus circuit depth, and fitting its mean versus m to an exponential $F_{\text{XEB}} = Af^m$, we can extract the polarization f, from which we can calculate r_Ω using Eq. (348).

The XEB protocol is defined for both Clifford [322] and non-Clifford [2,136,319–321] circuits. However, we note that the canonical circuits for XEB are the same circuits as in the "quantum supremacy" demonstrations, and XEB with Clifford circuits requires an impractical number of samples for low-uncertainty estimates of r_Ω (BiRB is a reliable alternative to XEB with Clifford circuits that has better statistical properties). XEB reliably estimates the average (in)fidelity of Ω-distributed layers under certain regularity conditions, including that the errors must be small [323]. As with other RB protocols, XEB is only a reliable, well-defined procedure if its success metric (F_{XEB}) decays exponentially. The theory of XEB shows that F_{XEB} will be an exponential (assuming small Markovian errors), but only for XEB circuits that are deeper than some minimal depth m_{min} [301,320–322]. This minimal depth is related to the scrambling rate of the Ω-distributed circuits that are chosen—i.e., how many Ω-random layers are needed to approximately transform any error map into an n-qubit depolarizing channel (note that the above theory simply assumes that each error map can be represented by such an n-qubit depolarizing channel). This minimum depth, therefore, depends on Ω and the layer set that Ω samples from (and therefore also on a device's connectivity) [301,320–322].

In addition to benchmarking the average infidelity of random n-qubit circuit layers, XEB can be structured to benchmark individual gates, layers of gates, or subcircuits that are fully scrambling. Figure 29 illustrates XEB performed on a two-qubit CZ gate. The benchmarked layers are composite layers consisting of (i) a layer of Haar random single-qubit gates on each qubit, and then (ii) a two-qubit CZ gate. So, each random layer is a "dressed" CZ gate. From the results in Fig. 29(b), one can extract a dressed process fidelity of $F_e = 98.2\%$. It should be noted, however, that unlike other methods for estimating individual gate (in)fidelities, such as interleaved RB (Sec. VIII F) and cycle benchmarking (Sec. VIII G), it is not as straightforward to separate the infidelity of the Haar random twirling gates from the infidelity of the interleaved gate [324]. Thus, by default, XEB always returns an estimate of the infidelity of the *dressed* gate or layer. However, one of the utilities of XEB is that it does not require the interleaved gate or layer to be Clifford (unlike interleaved RB or cycle benchmarking), and has been used to benchmark the fidelity of multiqubit non-Clifford gates, such as an *i*Toffoli [263], controlled-controlled-Z (CCZ) [218], and CCCZ gate [272]. As outlined in Appendix D 3 c, XEB can be extended to benchmarking qudit gates as well.

D. RB of general groups

Standard RB can benchmark any gate set that is both a group and a unitary 2-design (e.g., the 24 single-qubit Clifford gates), and modern native gate RB methods can directly benchmark a gate set that simply *generates* a group that is a 2-design (e.g., $\{X_{\pi/2}, Y_{\pi/2}\}$). However, some interesting gate sets generate (or are) groups that are not unitary 2-designs. For example, the CNOT, Hadamard, and Z gates generate the "real Clifford group," which is not a unitary 2-design [325]. Gate sets like this cannot be benchmarked either indirectly by standard RB or directly by (existing) native gate RB methods. Here, we discuss RB techniques that address this problem, and enable RB of gate sets that are groups but not unitary 2-designs [301,325–329].

The random circuits of standard RB can be constructed for any gate set that is a group, but when that group is not a unitary 2-design the average success probability $\bar{p}(m)$ of these circuits will not generally follow the simple exponential form $\bar{p}(m) = Af^m + B$, even approximately. Instead, the theory of twirling over general groups (see Sec. VIII A) implies that $\bar{p}(m)$ will be a sum over multiple exponential decays, and those exponential decays can be *matrix* exponentials. Specifically,

$$\bar{p}(m) \approx \sum_{\lambda=1}^{k} \text{Tr}(A_\lambda M_\lambda^m), \tag{367}$$

where the M_λ matrices contain average gate error information (i.e., together they can be used to compute the mean infidelity of the gates), and the A_λ matrices absorb all SPAM errors [301]. The exact functional form is determined by how the superoperator representation of a gate G decomposes into irreducible representations (see Sec. VIII A) of G. Each term in Eq. (368) corresponds to an irreducible representation in the decomposition of the superoperator representation, and the dimensions of M_λ and A_λ depend on the multiplicity of the corresponding irreducible representation. Reliably fitting data to multiexponentials is challenging [301], and it contrasts with the conceptual and practical simplicity of RB.

The literature on RB of groups that are not unitary 2-designs [301,325–330] is about creating (1) techniques for reliably analyzing data of the form given in Eq. (368) and/or (2) techniques for adapting the RB circuits and data analysis so that it is possible to separate out the multiexponential decay of Eq. (368) into individual exponential decays that can be separately analyzed. There are a variety of protocols for RB of particular groups that are not unitary 2-designs, including dihedral RB [327], real RB [325,326], and filtered RB [301,321]. Perhaps the most well-known RB protocol for general groups is *character RB* [212,301,328], and this is the only such protocol we discuss further.

1. Character RB

Character RB [212,301,328] is a conceptually elegant technique for RB of general groups that is also of practical importance (but which also has some important limitations). Character RB uses techniques from group representation theory to robustly isolate individual exponential decays in the multiexponential of Eq. (368). The general and somewhat abstract ideas underpinning character RB enable many practical RB protocols, including an RB technique designed for biased-noise qubits [331]. Many other RB or RB-adjacent protocols, such as simultaneous RB (Sec. VIII E) and cycle benchmarking (Sec. VIII G), use the same technique. A character RB experiment is determined by the following:

(a) a *benchmarking group* $\mathcal{G}$, which is the set of gates one aims to benchmark, and
(b) a *character group* $\hat{\mathcal{G}} \subseteq \mathcal{G}$, which is used to extract individual exponential decays robustly.

When run without errors, each character RB circuit implements a uniformly random element $\hat{G}$ of $\hat{\mathcal{G}}$. The results of running different circuits are added together with weights determined by $\hat{G}$ and a character function, which depends on the structure of $\hat{\mathcal{G}}$ and the decay the experiment aims to isolate.

Character RB is not capable of isolating each exponential decay for every group. If the benchmarking group is not multiplicity-free (i.e., the superoperator representation of the group contains multiple copies of one or more irreducible representations) the results of character RB will still include multiexponential decays [328]. Furthermore, character RB requires an impractical number of samples for some groups—a limitation that is addressed by filtered RB [301,321]. See Ref. [301] for a detailed discussion of RB of general groups.

E. Simultaneous RB

Simultaneous randomized benchmarking (sRB) is a widely used method for quantifying crosstalk errors [332]. It is perhaps the simplest of a variety of "advanced" RB techniques (discussed in Sec. VIII E–VIII G) that build on or expand standard RB (Sec. VIII B) and/or the native gate RB protocols (Sec. VIII C). These advanced RB methods are designed to measure gate set properties beyond the AGSI that those foundational RB techniques target.

The original sRB protocol consists of running single-qubit CRB on a qubit while either (i) idling neighboring qubits [Fig. 25(a)], or (ii) driving those qubits by independently running CRB in parallel on those qubits [Fig. 25(b)] [332]. These two *isolated* and *simultaneous* RB experiments result in two decay parameters (f_{iso} and f_{sim}) and corresponding error rates (r_{iso} and r_{sim}). Comparing these error rates quantifies the change in a qubit's gate error rate caused by driving neighboring qubits. Typically, $r_{\text{sim}} > r_{\text{iso}}$ due to crosstalk errors. The size of these crosstalk errors is sometimes quantified with the sRB number [332],

$$r_{\text{sRB}} = \frac{d-1}{d}(1 - f_{\text{sim}}/f_{\text{iso}}) \approx r_{\text{sim}} - r_{\text{iso}}\,. \tag{368}$$

Figure 30 shows data from running single-qubit CRB on two superconducting qubits while idling the other, as well as data from running CRB in parallel on the two qubits. We observe significant differences in the exponential decay rates between the isolated and parallel contexts, indicating that the single-qubit EPC is higher when gates are performed in parallel than in isolation. In this scenario—and in other superconducting qubit systems in general—the primary contributor to r_{sRB} is likely crosstalk-induced coherent errors acting on both qubits when they are operated simultaneously.

Running sRB on all the qubits in an n-qubit system requires $n+1$ different RB experiments. So, it is now common to run only the simultaneous RB experiment (and to still refer to this as "sRB"), measuring only r_{sim} for each qubit. In a many-qubit processor, those error rates (one for each qubit) quantity the infidelity of each qubit's gates when running single-qubit gates in parallel on every qubit. Dividing this into contributions from local and crosstalk errors for every qubit requires n more RB experiments, and is not necessary if the goal is to quantify the performance of many-qubit circuits. Therefore, these extra experiments are often skipped.

sRB can also be generalized to quantify crosstalk induced on or by multiqubit gates by running $n \geq 2$ qubit RB on a set of qubits while either idling all other qubits or running RB on those other qubits [314,333]. There are many ways to do this; for example, running single-qubit RB on all qubits quantifies simultaneous single-qubit crosstalk, running simultaneous two-qubit RB quantifies crosstalk between simultaneous two-qubit gates, or mixing single- and two-qubit RB captures crosstalk between simultaneous single- and two-qubit gates. Each choice for the parallel context will quantify a different aspect

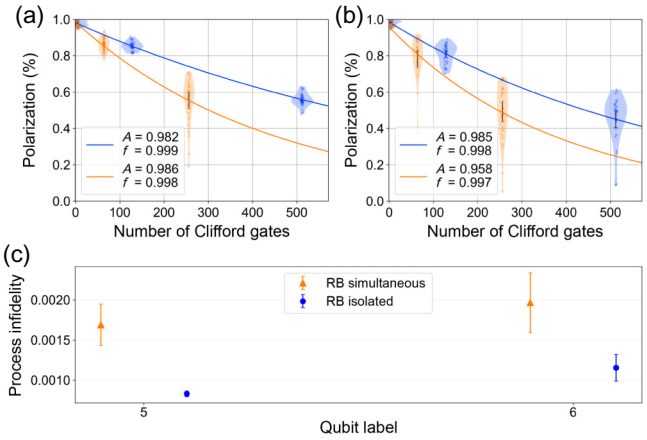


FIG. 30. Simultaneous RB. A demonstration of sRB, whereby RB is run in isolation and in parallel on two superconducting qubits labeled 5 and 6. (a),(b) Exponential decays for qubits 5 and 6, respectively, when running RB in isolation (blue) and simultaneously (orange). (c) The process infidelity (i.e., EPC) for each qubit measured in isolation ($e_{F,\mathrm{iso}}$, blue) and simultaneously ($e_{F,\mathrm{sim}}$, orange). The difference between $e_{F,\mathrm{iso}}$ and $e_{F,\mathrm{sim}}$ can be used to quantify the crosstalk errors induced by driving the other qubit.

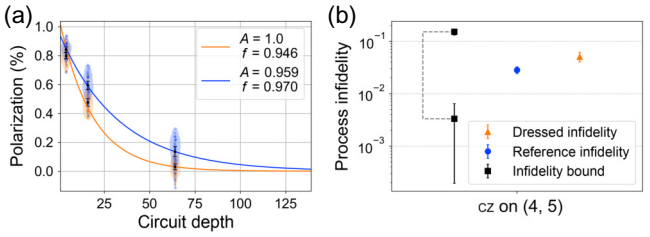


FIG. 31. Interleaved RB. IRB of a CZ gate between two superconducting qubits. (a) The *reference* (blue) and *interleaved* (orange) CRB decays. In the reference experiment, circuit depth (m) is the number of uniformly random two-qubit Clifford gates. In the interleaved experiment, each of the random Clifford gates is followed by a CZ gate, resulting in longer circuits and typically a lower average success probability at the same m value. (b) The estimated process infidelity extracted from the reference (blue, e_F) and interleaved (orange, $e_{F,D}$) experiments. We estimate the CZ gate's process infidelity [see Eq. (372)] to be $e_{F,\mathrm{CZ}} = 2.3(6)\%$. IRB is intended to estimate a gate G's infidelity ϵ_G, but the IRB error rate can be very different from ϵ_G due to systematic flaws in IRB. Upper and lower bounds [computed from Eq. (374)] on CZ's infidelity are shown in black. The error bars on this region are computed from the statistical uncertainties in the estimates of e_F and $e_{F,D}$.

of device crosstalk. Notably, a version of simultaneous two-qubit RB is used to define the *error per layered gate (EPLG)* [314], which quantifies the error per qubit when a random layer containing one- and two-qubit gates is applied to all qubits in a system simultaneously. As of 2025, IBM reports the EPLG as one of its primary performance metrics for its cloud-access systems.

Implementing sRB requires addressing a scheduling problem that becomes worse as n increases [334]. This is because CRB's random n-qubit Clifford gates get compiled into circuits of native gates of varying lengths (with typical depth increasing with n). This problem can be avoided if sRB does not use CRB, but instead uses a native gate RB protocol (Sec. VIII C), such as DRB, BiRB, or XEB [314]. Finally, we note that data from sRB experiments can also be used to learn more than just r_{sRB}, with the aid of a variety of more complex methods [334–336] that enable estimating, for example, the rates of correlated errors between different pairs of qubits.

F. Interleaved RB

Interleaved randomized benchmarking (IRB) [184] is a method for estimating the infidelity of an individual Clifford gate (extensions to some non-Clifford gates exist [337,338]). IRB is typically used to estimate the infidelity of a one- or two-qubit gate, but in principle it can be applied to n-qubit gates for any n (e.g., a many-qubit layer of parallel one- and two-qubit gates). IRB for an n-qubit gate G is a simple extension of n-qubit CRB. It consists of two independent RB experiments. One experiment is often called the *reference* RB experiment and it consists of simply running standard CRB to estimate the CRB decay parameter (f) and the corresponding EPC (r). The other experiment—the *interleaved* RB experiment—consists of again implementing CRB, but now each randomly sampled Clifford gate is followed by G [see Fig. 25(d)], i.e., a depth-m interleaved circuit has the form

$$\mathcal{C}_{m,G} = C_{m+2} \circ G \circ C_{m+1} \cdots \circ G \circ C_2 \circ G \circ C_1 \,, \quad (369)$$

where $C_1, C_2, \ldots, C_{m+1}$ are independent and uniformly sampled Clifford gates (as in CRB), and C_{m+2} is the unique Clifford gate that inverts the entire preceding sequence. The interleaved RB experiment also produces an estimated decay parameter (f_D) and corresponding error rate (r_D). Figure 31(a) shows reference and interleaved RB decay curves for IRB of a CZ gate between two superconducting qubits. The interleaved curve decays faster than the reference curve due to the additional gate G inserted at each circuit depth.

The error rate r_D is an estimate of the mean infidelity of G composed with (i.e., "dressed" by) a uniformly random Clifford gate, *not* an estimate of G's infidelity. The standard IRB analysis attempts to subtract the contribution of the uniformly random Clifford gate's error to r_D, by comparing r_D to r. Specifically, IRB's estimate of the gate G's *average gate infidelity* is defined by

$$r_G = \frac{d-1}{d}\left(1 - \frac{f}{f_D}\right) \approx r_D - r \,. \quad (370)$$

Alternatively, IRB's estimate of the gate G's *process infidelity* is given by

$$e_{F,G} = \frac{d^2-1}{d^2}\left(1-\frac{f}{f_D}\right) \approx e_{F,D} - e_F\,. \tag{371}$$

Applying Eq. (372) to our CZ gate data in Fig. 31, we estimate CZ's process infidelity to be $e_{F,\mathrm{CZ}} = 2.3(6)\%$ [see Fig. 31(b)].

It is important to highlight that IRB is not generally a reliable method for estimating a gate's infidelity. This is primarily because unitary errors in G can coherently add or cancel with errors in the random Clifford gates C_i, and this can even cause the interleaved curve to decay more slowly than the reference curve—resulting in a *negative* IRB error rate!—even when G's errors are large. This implies that there is a systematic and potentially large discrepancy between r_G and G's true infidelity, ϵ_G (we call these discrepancies *systematic* as they are *not* due to shot noise; i.e., they are not statistical in origin). For the average gate infidelity, IRB theory shows that r_G and ϵ_G are related by the inequalities

$$\epsilon_G - E < r_G < \epsilon_G + E\,, \tag{372}$$

where

$$E = \min\left\{\begin{array}{l}\dfrac{(d-1)}{d}\left[\left|f-\dfrac{f_D}{f}\right|+(1-f)\right] \\ \dfrac{2(d^2-1)(1-f)}{d^2 f} + \dfrac{4\sqrt{1-f}\sqrt{d^2-1}}{f}\end{array}\right\}, \tag{373}$$

and $d = 2^n$. The upper and lower bounds in Eq. (373) can span orders of magnitude, and a tighter relationship between r_G and ϵ_G can only be guaranteed if more is known about the errors—e.g., if it is known that coherent errors form a small contribution to infidelity [339]; see Sec. VIII H 1. In Fig. 31(b), we plot these upper and lower bounds on the estimated process infidelity for the CZ gate. When this systematic error is combined with the statistically uncertainties in our estimates of f and f_D, this range spans over two orders of magnitude, ranging from above 10^{-1} to below 10^{-3}.

IRB has been widely used, but its large systematic errors have caused it to become less popular in recent years. There are now a variety of alternatives to IRB, including many RB or RB-like techniques for measuring gate infidelities (as well as SPAM-error-robust tomographic techniques like gate set tomography; see Sec. VII D). Many of these techniques are also more scalable than IRB (IRB inherits the scaling problems of CRB, discussed in Sec. VIII C). One such technique is cycle benchmarking, which we discuss in detail in Sec. VIII G. Other examples include fitting error models directly to RB data [311,340]; running native gate RB protocols with different sampling distributions Ω and using simple linear algebra to estimate different gates' infidelities [310,311]; interleaved versions of character RB (Sec. VIII D 1), which is closely related to cycle benchmarking; and Pauli noise learning techniques (see Sec. IX C).

G. Cycle benchmarking

Cycle benchmarking (CB) [341] is a protocol for estimating the process fidelity of an n-qubit gate, a.k.a. a "layer" or "cycle." Following CB convention, here we will use the "cycle" terminology, which is defined to be a set of gates acting on disjoint sets of qubits all occurring during the same moment in time, in analogy with a clock cycle on classical computers. CB is an alternative to IRB

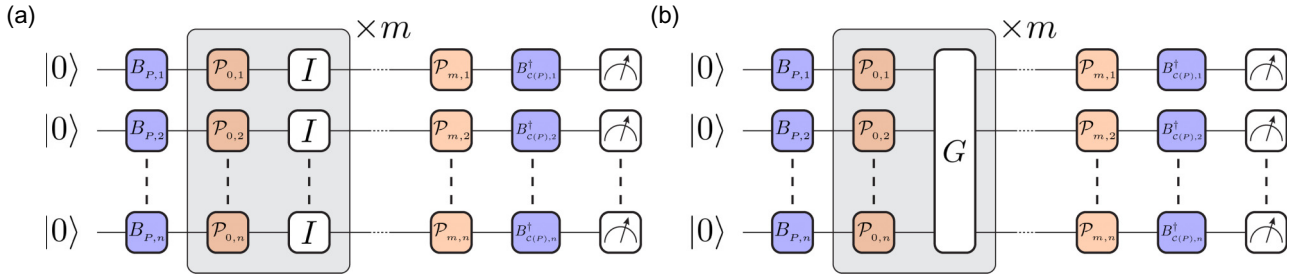


FIG. 32. Cycle benchmarking circuits. (a) CB of the all-identity "reference" cycle. $B_{P,q}$ denotes the basis preparation gate on qubit q for a random eigenstate of the Pauli P, $\mathcal{P}_{i,q}$ denotes the ith twirling operator acting on qubit q, and $B^\dagger_{C(P),q}$ rotates qubit q back to the initial eigenstate of P at the end of the circuit. Explicit identity gates I have been inserted for visual clarity, but this cycle is either skipped in compilation, or the identity gates can be implemented as true idles for the duration of a cycle of single-qubit or two-qubit gates; the choice is up to the experimenter. If the identity gates are skipped in compilation, then this measures the average fidelity of a cycle of random Pauli gates applied simultaneously to all n qubits. (b) CB of an n-qubit gate cycle G. G can be composed of any combination of single- and multiqubit gates, as long as $G^m = \mathbb{I}$ for a sequence depth of m.

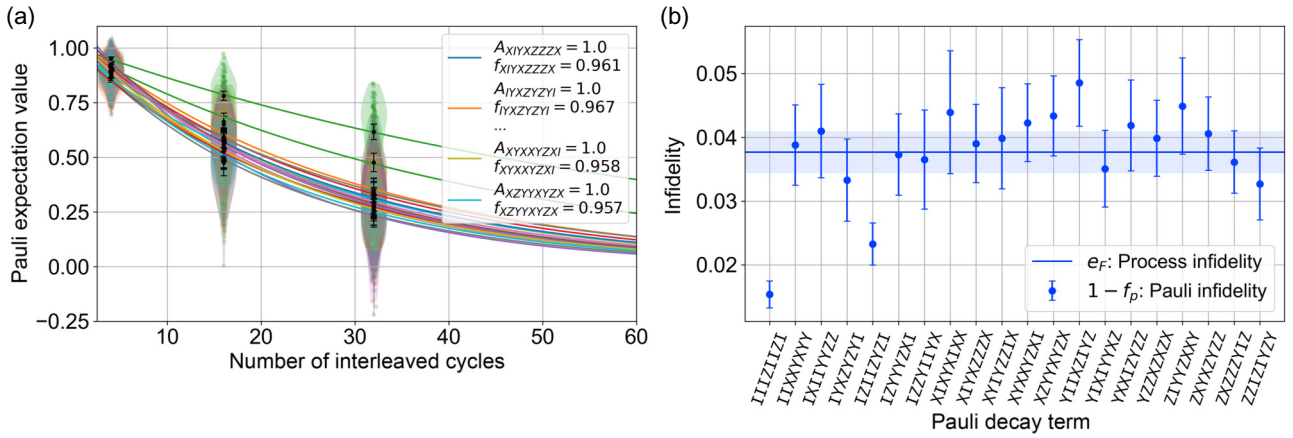


FIG. 33. Cycle benchmarking of simultaneous single-qubit Paulis. (a) Pauli decays for each Pauli basis P, with the SPAM parameter A_P and the exponential fit parameter f_P listed in the legend for a subset of P. (b) Pauli infidelities $e_P = 1 - f_P$ for each Pauli P, and the average process infidelity e_F (horizontal blue line). The highlighted region denotes the 95% confidence interval of e_F.

that is arguably more useful in practice. CB interleaves the cycle of interest in between layers of random Pauli gates (see Fig. 32), instead of the layers of random n-qubit Clifford gates used in IRB. The Pauli group implements a weaker twirl than the Clifford group—it converts a general error map to a Pauli stochastic channel, rather than a global depolarizing channel (see Appendix C 4). But Pauli twirling requires only a single layer of parallel single-qubit Pauli gates, whereas n-qubit Clifford gates require many one- and two-qubit gates. This makes CB much more scalable than IRB, so CB enables benchmarking many-qubit cycles containing parallel one- and two-qubit gates.

CB estimates the process fidelity of a cycle of gates, and we describe CB for the case of an n-qubit cycle containing only Clifford gates:

(1) For K different n-qubit Pauli operators P, that are uniformly sampled if $n \gg 1$ but can consist of every possible Pauli operator if n is small:

   (a) Use a layer of single-qubit gates to prepare the qubits in a random tensor-product eigenstate of P.
   (b) Apply a circuit consisting of m applications of the cycle of interest G interleaved with cycles of randomly sampled n-qubit Pauli operators (applied via randomized compiling), for a range of values m that all satisfy $G^m = \mathbb{I}$.
   (c) Measure the Pauli operator P, whose estimated value we denote by $f_{P,m}$, which is typically achieved using a layer of single-qubit gates and a computational basis measurement.
   (d) Fit $f_{P,m}$ to an exponential decay of the form

$$f_{P,m} = A f_P^{\,m}\,, \tag{374}$$

where A absorbs all SPAM errors. The fit value for f_P is an estimate of

$$f_P = \left(\prod_{k=1}^{j} \lambda_{G^k P G^{-k}}\right)^{1/j}, \tag{375}$$

where j is the smallest integer such that $G^j = \mathbb{I}$ and $\lambda_P = \Lambda_{PP}$, where Λ is G's error channel's PTM and Λ_{PP} is the diagonal element indexed by Pauli operator P.

(e) Estimate the process fidelity of the cycle to be

$$F_e = \frac{1}{K}\sum_{i=1}^{K} f_{P_i}\,. \tag{376}$$

Individual exponential decays in CB are often referred to as *Pauli decays* and are labeled by the Pauli operator P specifying the basis of the state preparation and measurement. If CB is applied to the idle (i.e., identity) cycle (see Fig. 33), each Pauli decay curve measures the eigenvalues of the PTM of G, but for more general cycles some of the f_P correspond to products of eigenvalues of G's error channel (see Appendix E). This complication is encompassed by Eq. (376). As a result, F_e is not an accurate estimate of the process fidelity in general. However, it is proven in Ref. [341] that F_e is a lower bound on the true process fidelity of the cycle (in the limit of infinite samples). The number of Pauli decays required to obtain a fixed estimation precision is independent of the number of qubits, and instead only depends on the infidelity of the cycle, which follows from standard statistical analysis of RB protocols [290,341]. As a general guide, a minimum of $K = \min(20, 4^n - 1)$ Pauli operators should be sampled from $\mathbb{P}_n$ for low uncertainty estimates of the process fidelity [341,342].

CB measures the process (in)fidelity of a *dressed* cycle. Therefore, the error rate measured by CB contains contributions from both errors in the interleaved cycle and the random Pauli gates. This is the relevant error rate for cycles that will be used in randomly compiled or Pauli frame randomized circuits [155]. But, it is also possible to approximately isolate the process infidelity of the "bare" interleaved cycle (G), by implementing CB with (i) the cycle of interest and (ii) a reference cycle containing no gates [see Fig. 32(a)], and then applying exactly the same analysis as in IRB [see Eq. (372)]. This has the same fundamental limitations as IRB (see the discussion in Sec. VIII F), but in practice the systematic error in this estimation method is typically significantly smaller than in IRB. This is because the fidelity of a random Pauli gate (the randomizing gates in CB) is typically higher than the fidelity of a random Clifford gate (the randomizing gates in CRB). For example, Ref. [343] used CB to estimate the fidelity of a CZ gate and [using Eq. (374)] found lower and upper bounds on its fidelity of 97.52(2)% and 99.764(5)%, respectively, whereas when using IRB these lower and upper bounds were 91.9(2)% and 99.96(1)%, respectively.

CB is simplest and most efficient for benchmarking Clifford cycles, but it can also be used to benchmark non-Clifford gates. Doing so requires adding correction gates to the end of CB circuits to return the qubits to a Pauli eigenstate. This can require many multiqubit gates at the end of each benchmarking circuit. Therefore, to reliably benchmark non-Clifford gates, the native gates required to implement the correction operations (which usually include the interleaved gate itself) must have high enough fidelity that the correction operations do not corrupt the measured fidelity of the dressed cycle. For example, Ref. [344] benchmarked non-Clifford $\mathrm{CS} = \sqrt{\mathrm{CZ}}$ and $\mathrm{CS}^\dagger$ gates. CB was also used to benchmark a three-qubit non-Clifford *i*Toffoli gate [263], which would not have been feasible using three-qubit non-Clifford RB.

One utility of CB is that it can benchmark the process fidelity of an entire cycle of gates containing any combination of single- and multiqubit gates (similar to methods like MRB), as long as the cycle composes to the identity operation at some circuit depth m [345]. Thus, CB can *holistically* quantify the impact of crosstalk between gates in a parallel gate cycle. For example, it can be used to measure crosstalk experienced by idling spectator qubits during a two-qubit gate. To demonstrate this, in Fig. 34 we plot the process infidelity of eight different CZ gates measured via CB on an eight-qubit superconducting quantum processor with a ring topology. Additionally, we measure the process infidelity of cycles containing each of the eight CZ gates, as well as idle gates on the spectator qubit on either side of the CZ gate (i.e., the interleaved gate cycle is $G = I \otimes \mathrm{CZ} \otimes I$). We observe that, in all cases, the cycle with the idle qubits has a larger process infidelity than the cycle containing just the CZ gates. This highlights two important concepts: (i) it should not be assumed that gates have no impact on idle qubits (and vice versa [346]), and (ii) when understanding circuit performance, it is most informative to benchmark the constituent cycles as they appear in the circuit.

H. Purity benchmarking

Purity benchmarking (PB) [347–349] is a family of RB techniques for quantifying how coherent a gate set's errors are. Purity benchmarks provide complementary information to the foundational RB protocols (i.e., the group and native-gate RB protocols), which intentionally mix together all kinds of errors into a single error rate. As discussed in Sec. III, Markovian errors can be broadly categorized as either coherent (i.e., unitary) or incoherent (i.e.,

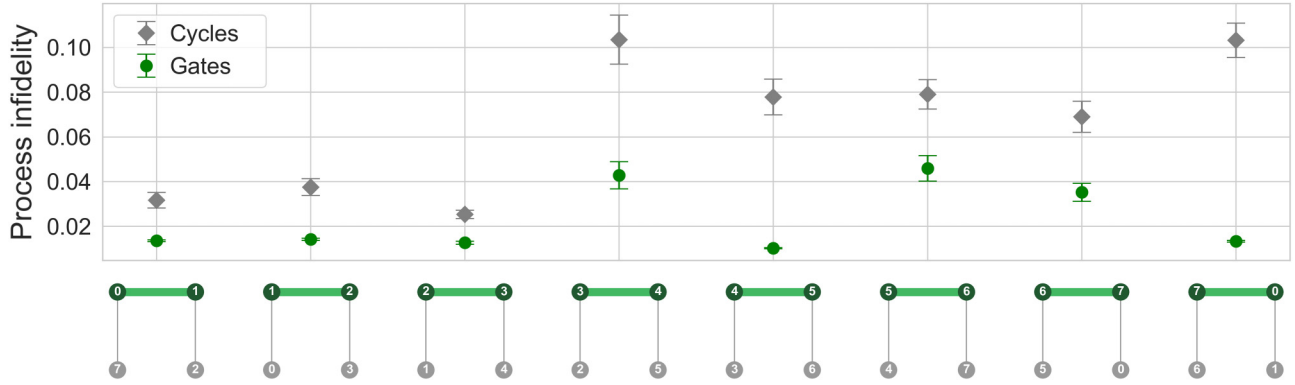


FIG. 34. CB of gates vs. cycles. The process infidelities of eight different two-qubit CZ gates measured via CB are plotted in green. When the two nearest-neighbor idling qubits are benchmarked alongside each CZ gate (gray), the CB performance is worse in all cases. Notably, the good performance of an isolated CZ gate does not guarantee the good performance of the cycle, which includes idling spectator qubits. For example, the CZ between qubits 4 and 5 has the lowest individual process infidelity, but has one of the worst process infidelities when the idles are included.

stochastic). PB methods can be used to quantify the relative contributions of coherent and stochastic errors, and they are based on the purity γ of a quantum state ρ. The purity of a quantum state is

$$\gamma = \mathrm{Tr}(\rho^2) = \frac{1}{d}\left(1 + \|\mathbf{r}(\rho)\|^2\right), \tag{377}$$

where

$$\rho = \frac{1}{d}[\mathbb{I} + \mathbf{r}(\rho)\cdot\boldsymbol{\sigma}], \tag{378}$$

$\mathbf{r}(\rho)$ is the generalized d-dimensional Bloch vector, and $||\mathbf{r}(\rho)||^2$ is its Euclidean norm (i.e., average squared length). Here $\boldsymbol{\sigma}$ is the vector of Pauli matrices.

In the context of randomized benchmarks, we prepare quantum states using sequences of random gates, and the final state $\rho' = \mathcal{E}(\rho)$ will have a purity $\gamma \leq 1$, which depends on the nature of the error channel $\mathcal{E}$ (e.g., $\mathcal{E}$ can be some mixture of coherent and stochastic errors). One way to quantify how coherent the error channel $\mathcal{E}$ is in terms of *unitarity* of $\mathcal{E}$,

$$u(\mathcal{E}) = \frac{1}{d-1}\int d\psi \, \|\mathbf{r}[\mathcal{E}(|\psi\rangle\langle\psi|)] - \mathbf{r}[\mathcal{E}(\mathbb{I}/d)]\|^2, \tag{379}$$

which is the Euclidean norm of the Bloch vector of the state $\mathcal{E}(|\psi\rangle\langle\psi|)$ (with the identity component subtracted off), averaged over all pure states. If $\mathcal{E}$ is a unitary channel, then $u(\mathcal{E}) = 1$, and $u(\mathcal{E}) < 1$ if $\mathcal{E}$ includes contributions from stochastic noise. While not all purity benchmarks utilize the unitarity, Eq. (380) demonstrates that it is possible to *quantify* the relative contributions of stochastic and coherent errors to the AGSI of an RB experiment. In this subsection, we discuss several different randomized benchmarks, which attempt to quantify the relative error rates of coherent and stochastic errors in a gate set. While this goes beyond the scope of this tutorial, it should be noted that gate set tomography (Sec. VII D) can also be used to quantify the amount of coherent errors and stochastic noise in a gate [46].

1. eXtended RB

eXtended randomized benchmarking (XRB) [342,347] is a PB protocol that is based on CRB. XRB estimates the average unitarity of a set of n-qubit Clifford gates. XRB requires only a small modification to the standard CRB protocol: XRB performs the standard CRB circuits introduced in Sec. VIII B, but instead of performing an inverting operation at the end of the sequence, state tomography is performed on the resulting state in order to estimate the length of the Bloch vector. XRB characterizes the unitarity in terms of the decay rate of the average squared Bloch vector length with sequence depth. For a single qubit, the average squared Bloch vector length,

$$\tilde{\gamma} = ||\boldsymbol{r}(\rho)||^2, \tag{380}$$

is equivalent to

$$\tilde{\gamma} = \langle X\rangle^2 + \langle Y\rangle^2 + \langle Z\rangle^2. \tag{381}$$

This is a shifted and rescaled version of Eq. (378).

XRB consists of (i) running CRB circuits for various depths m without the inversion gate, (ii) estimating $\tilde{\gamma}$ for each circuit, and then (iii) fitting the mean of $\tilde{\gamma}$ (which we denote by $\langle\tilde{\gamma}(d)\rangle$) as a function of m, to

$$\langle\tilde{\gamma}(m)\rangle = Au^m. \tag{382}$$

The fit value for u is an estimate of the mean unitarity of the benchmarked gates. This can then be used to estimate the *stochastic process infidelity* $e_S(\mathcal{E})$ defined by

$$e_S = 1 - \sqrt{\frac{(d^2-1)u + 1}{d^2}}. \tag{383}$$

If standard CRB is also performed in addition to XRB, then the process infidelity e_F measured via CRB represents the total error. Together, one can estimate the *coherent process infidelity* e_U by $e_U = e_F - e_S$.

In Fig. 35(a), we plot exponential decays for CRB and XRB. In Fig. 35(b), we compare the process infidelity of the gates e_F (measured via CRB) with the estimated stochastic process infidelity e_S (measured via XRB); the difference between the two is the coherent process infidelity e_U. The stochastic process infidelity is an approximate measure for determining whether or not a gate set is *coherence limited* (i.e., all gate errors are due to incoherent noise). If $e_F = e_S$, then $e_U = 0$ and, thus, the gates have no coherent errors. However, because infidelity is only sensitive to coherent errors at $\mathcal{O}(\theta^2)$, if an estimate of e_S is equal to e_F within error bars, it is possible that coherent errors still exist—XRB does not amplify coherent errors, so it is an inefficient method for estimating the size of coherent errors.

One application of XRB is to quantify the magnitude of crosstalk errors. In Fig. 35(c), we show the CRB and XRB process infidelities for single-qubit gates performed in isolation and simultaneously for two qubits. We see that the CRB infidelity (e_F) for each qubit is larger for simultaneous CRB than for isolated CRB, but the XRB process infidelity (e_S) is approximately the same in both cases. This demonstrates that crosstalk-induced coherent errors make up a larger fraction of the total error rate under simultaneous operation.

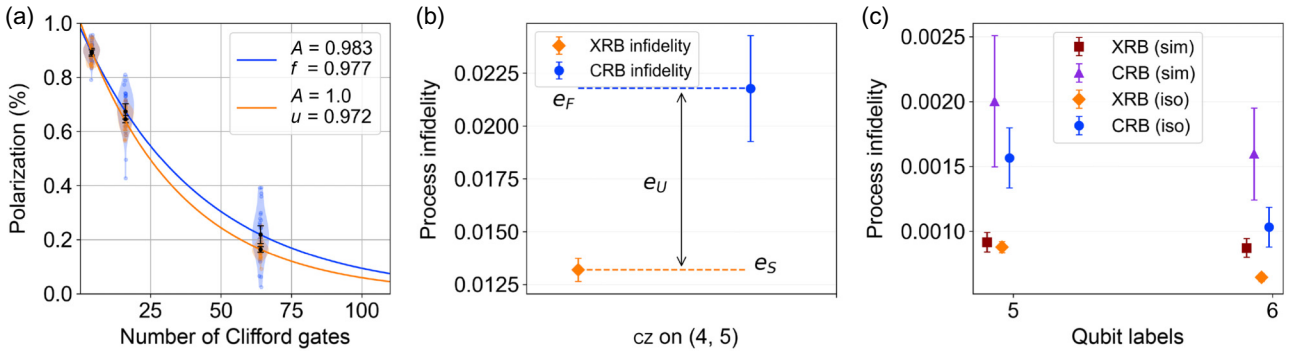


FIG. 35. eXtended RB. (a) Exponential decays for two-qubit CRB (blue) and XRB (orange). All two-qubit Cliffords are decomposed into native single-qubit gates and a native two-qubit CZ gate. (b) CRB (blue) and XRB (orange) process infidelities for the exponential decays in (a). The CRB process infidelity e_F is the AGSI for two-qubit Cliffords, and the stochastic process infidelity e_S is the approximate coherence limit for two-qubit Clifford gates. The difference between the two $e_U = e_F - e_S$ is the average error rate due to coherent errors. (c) Isolated vs. simultaneous single-qubit CRB and XRB. The CRB process infidelity is larger for both qubits under simultaneous operation, whereas the stochastic process infidelities are approximately equal in both cases, indicating the presence of coherent crosstalk errors between simultaneous single-qubit gates.

2. Speckle purity benchmarking

Speckle purity benchmarking (SPB) [2] is a protocol for estimating the decay of the purity of states produced by random circuits versus circuit depth. This method is based on the observation that, for long enough circuits, the probability $p(x)$ of observing a particular bit string x will be a random variable with a Porter-Thomas distribution. For most circuits, $p(x)$ will be exponentially close to zero. But for some rare circuits x will appear with significantly higher probabilities. The observed data will then demonstrate a "speckle pattern" when presented visually. However, when depolarizing errors dominate, the speckle pattern will be smoothed out as the distribution approaches the uniform distribution [see Figs. 29(c) and 29(d)].

SPB is the following procedure: sample N random scrambling circuits of depth m, i.e., circuits with properties similar to those typically used in XEB. Now, choose some bit string x, and let $\mathcal{P}$ be the probability of measuring bit string x *assuming the circuits are run without errors*. Because the circuit is random, $\mathcal{P}$ is a random variable. If the gates are unitary, then for sufficiently deep circuits, $\mathcal{P}$ will be distributed according to the Porter-Thomas (PT) distribution, whose probability density is:

$$f_{\mathrm{PT}}(p) = (d-1)(1-p)^{d-2}\,, \tag{384}$$

with variance

$$\sigma^2_{\mathrm{PT}} = \frac{d-1}{d^2(d+1)}\,. \tag{385}$$

On the other hand, if the gates completely depolarize the state, then all output strings become equally probable. In this case, we can again describe the probability as a random variable, but with a trivial ($\mathcal{T}$) probability density function:

$$f_{\mathcal{T}}(p) = \delta\left(p - \frac{1}{2^d}\right)\,, \tag{386}$$

with δ the Dirac delta distribution, and the variance of $\mathcal{T}$ is zero

$$\sigma^2_{\mathcal{T}} = 0\,. \tag{387}$$

In real experiments, there are stochastic and coherent errors, and the stochastic errors push $\mathcal{P}$ towards the trivial distribution with increasing circuit depth, whereas coherent errors preserve the PT distribution. In particular, the probability $\mathcal{P}$ of measuring a given bit string in a depth m circuit will be approximately given by a mixture of the PT distribution and the trivial distribution:

$$f_{\mathrm{exp}}(p) \approx (1-\epsilon)^m f_{\mathrm{PT}}(p) + \epsilon^m f_{\mathcal{T}}(p)\,, \tag{388}$$

with ϵ the rate of stochastic errors per circuit layer (more precisely, $1-\epsilon$ is the process polarization corresponding to the stochastic portion of the error channels, and so ϵ is approximately the rate of stochastic errors except for very few qubits). SPB theory relates the variance of this distribution to the variance of the PT distribution as

$$\sigma^2_{\mathrm{exp}} = \epsilon^{2m}\sigma^2_{\mathrm{PT}}\,. \tag{389}$$

Therefore, we can estimate the purity decay with cycle depth by simply observing the rate at which the variance of the distribution of bit-string probabilities decays. For example, by comparing the XEB fidelity decay and purity decay for the data in Fig. 29(b), we can investigate the relative size of coherent and incoherent errors in the system. Using SPB, we estimate that roughly 68% of the dressed CZ's error can be attributed to stochastic errors; the remaining are attributed to coherent errors.

3. Iterative RB

Iterative RB protocols are RB-like methods for amplifying gate errors so that it is possible to separate coherent errors from stochastic noise. These methods depend on the fact that constructively interfering coherent errors will grow quadratically with circuit depth. To see this, consider the simple example of applying many $R_x(2\pi)$ rotations to a qubit initially in the ground state, but each time the qubit over-rotates by a small angle θ. The resulting state of the qubit after M rotations is

$$|\psi\rangle = \prod^{M} e^{-i\theta\sigma_x} |0\rangle = \cos(M\theta)\,|0\rangle - i\sin(M\theta)\,|1\rangle\ . \tag{390}$$

The fidelity of this state with respect to $|0\rangle$ is $F = |\langle 0|\psi\rangle|^2 = \cos^2(M\theta) \approx 1 - (M\theta)^2$, thus the infidelity is $\epsilon = 1 - F \approx (M\theta)^2$. Therefore, the infidelity scales quadratically in both the over-rotation angle θ and the number of rotations M. In contrast, stochastic errors typically only grow linearly with circuit depth: if we take p to be the probability of a stochastic error per gate, then $1-p$ is the probability of no error per gate, and $(1-p)^M$ is the probability of no error after M gates. For a circuit with M total gates, $\epsilon = 1 - (1-p)^M \approx Mp$ is the probability of an error after M gates. Thus, stochastic errors accumulate linearly with circuit depth in the small error limit.

Iterative RB [350] interleaves M repetitions of a target quantum gate within a standard CRB sequence (see Sec. VIII F), with M varied. Because the coherent errors in the gate will grow quadratically in M, one can fit the fidelity decay of the sequence to both quadratic and linear functions, with the quadratic component capturing the coherent contributions to the gate error, and the linear component capturing the incoherent contributions to the gate error. This method can be adapted to a variety of randomized benchmarks [351–353].

I. RB with non-Markovian errors

All of the RB protocols discussed so far in this section are primarily based on theory that assumes Markovian errors. Those RB protocols are therefore not guaranteed to work correctly in the presence of non-Markovian errors. In general, standard RB data is not guaranteed to follow a simple exponential decay in the presence of non-Markovian errors [196,198,309]. For example, $1/f$ noise is well-known to cause nonexponential RB decays [288]. Furthermore, the RB protocols discussed so far are not designed to learn anything about the rates of non-Markovian errors (although they will incorporate the rates of some kinds of non-Markovian errors into the measured RB error rates). There are, however, a variety of adaptations to RB protocols that enable learning about one or more kinds of non-Markovianity using RB. Examples include time-resolved RB [96] and loss RB [354], which measure drifting gate error rates versus time and qubit loss rates, respectively. Here, we discuss only the most-widely used RB protocols for non-Markovian errors: those designed for quantifying leakage.

1. Leakage RB

RB protocols that measure leakage rates are relatively simple to implement and are widely used. Leakage describes an error in which a qubit is excited out of the computational basis state to higher energy levels [see Fig. 4(f)]. This is a common source of error in systems whose energy spacings are not sufficiently well separated to isolate the $|0\rangle \longrightarrow |1\rangle$ transition from transitions to higher energy levels. Leakage cannot be captured by most RB protocols, and it can corrupt their results. However, many RB protocols can be modified to account for leakage [98,100,355]. There are a variety of ways to quantify leakage using RB methods, but the conceptually simplest methods consist of running standard RB experiments (or native gate RB experiments) while monitoring the $|2\rangle$ state population (or higher states). This method is often termed *leakage randomized benchmarking (LRB)*, and we focus our discussion on this simple technique.

LRB is the following simple adaptation to standard single-qubit CRB:

(1) Run standard CRB circuits and, at the end of each circuit, measure whether the final state of the qubit is $|0\rangle$, $|1\rangle$, or $|2\rangle$.
(2) Fit the average $|2\rangle$ state population versus circuit depth to a simple exponential growth function, to estimate the leakage rate per Clifford gate (r_l).

To perform LRB, it is therefore necessary to be able to readout the $|2\rangle$ state (although note that LRB is robust to errors in this readout). Figure 36 illustrates simultaneous LRB on two transmon qubits, and how it can be used to identify qubits with high leakage rates.

IX. PARTIAL TOMOGRAPHY AND FIDELITY ESTIMATION

QCVV methods can be categorized by (i) the amount (and type) of information they provide, and (ii) the cost—in both experimental and computational resources—to run them. Gaining more information typically requires more resources, and so QCVV methods can often be placed on a sliding scale from (1) highly informative but costly, to (2) highly efficient but providing little information [29]. Tomography of quantum states, processes, measurements, or gate sets (see Sec. VII) is at one extreme of this continuum: these methods provide comprehensive information about the models and rates of all possible kinds of (Markovian) errors, but they require resources that are exponential in system size. This limits the application of these methods to the few-qubit setting.

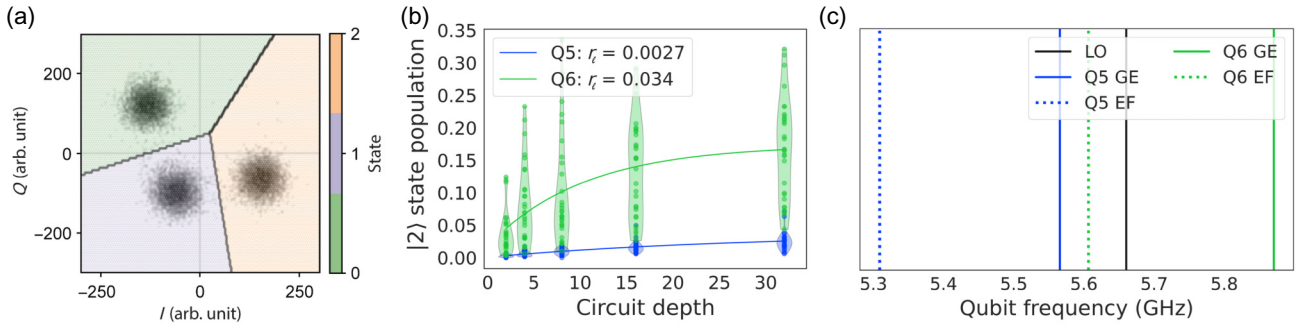


FIG. 36. Leakage RB. There are a variety of RB protocols that can quantify rates of leakage errors. This figure demonstrates one such method, which we call *leakage RB* (LRB), on transmon qubits. This method uses qutrit readout to discriminate between leaked and computational basis states. (a) Example of readout and classification boundaries for classifying $|0\rangle$, $|1\rangle$, or $|2\rangle$ for a superconducting qubit. (b) LRB data for two qubits (labeled Q5 and Q6) under simultaneous operation, showing the probability of observing the 2 outcome as a function of the RB circuit's depth. This data is fit to an exponential to estimate the leakage rate (r_l). Q6 has a much larger leakage rate ($r_l = 3.4\%$) and steady-state $|2\rangle$ state population than Q5 ($r_l = 0.27\%$). Leakage can be due to either coherent excitation to higher energy levels or thermal noise, but the frequency spectrum of the two qubits shown in (c) indicates that the $|0\rangle \to |1\rangle$ (i.e., "GE") transition frequency of Q5 is close to the $|1\rangle \to |2\rangle$ (i.e., "EF") transition frequency of Q6. Therefore, the leakage on Q6 is most likely due to crosstalk from Q5 when performing simultaneous single-qubit gates.

In contrast, most randomized benchmarks (see Sec. VIII) are extremely efficient to run but they provide only one (or a handful) of numbers summarizing a gate set's performance (e.g., the average fidelity of a gate set). It is, however, both possible and often useful to obtain more information about a system (e.g., a gate set) than provided by randomized benchmarks, without resorting to exponentially expensive tomography. In this section, we discuss techniques that sit in the middle of this cost versus information sliding scale, which generally fit into one (or both) of two categories: (1) "partial tomography" methods, and (2) "fidelity estimation" methods.

There are many more resource-efficient characterization and benchmarking techniques than can be covered here. Therefore, we list some important (and partially overlapping) categories of partial tomography and fidelity estimation, covering only some of them in detail below:

(a) *Targeted tomography*. An n-qubit state, gate, or gate set contains exponentially many (in n) independent parameters, so it is infeasible to learn all of them for $n \gg 1$. However, it is possible to learn a subset of those parameters, or some function of those parameters. There are a variety of partial tomography techniques that target specific parameters in a state, process, or gate set [356–363]. For example, there are techniques for learning one or more of a transfer matrix's eigenvalues, e.g., spectral tomography [362] and phase estimation (see Sec. VI E). One way to reduce the cost of process tomography is to instead learn its action as a classical, probabilistic gate. We discuss this method, sometimes called "truth table tomography," in Sec. IX A.

(b) *Randomized measurement methods*. A variety of partial tomography methods are built on the idea of measuring a quantum state or process in a small number of randomly chosen bases [364]. Methods of this sort include shadow tomography [210,365,366], direct fidelity estimation [367,368], and cycle benchmarking (which we covered in Sec. VIII G). We discuss direct fidelity estimation in Sec. IX B.

(c) *Pauli noise learning*. Stochastic Pauli channels (see Sec. III E) are a practically relevant class of error channels that contain only 4^n parameters, rather than the 16^n parameters of a general error map. There are a variety of techniques that (i) use twirling or randomized compiling [154,155] to enforce a stochastic Pauli noise model, and (ii) learn the parameters of those Pauli channels. Many of these methods have close connections to randomized benchmarks (Sec. VIII), and some of them can efficiently learn *sparse* Pauli channels that contain only a polynomial number of unknown parameters. We discuss some of these methods in Sec. IX C.

(d) *Ansatz tomography*. A range of tomographic methods exist that reconstruct states, processes, or gate sets more efficiently than the "brute force" methods discussed in Sec. VII by assuming or privileging some simplifying structure. Some of these methods *assume* a structure—e.g., that the state is pure or local, or that the process is unitary or Pauli stochastic—and will give an incorrect estimate if the assumption is violated. Better methods use a hierarchical *ansatz*—e.g., that the density matrix or process matrix has low rank—and perform efficiently when the ansatz is satisfied, yet also recognize

and characterize (less efficiently) objects that do not satisfy the ansatz. Examples include methods for tomography of pure states and unitaries [369, 370], states satisfying strong symmetries [356,357, 371,372], low-rank states and processes (including compressed sensing approaches) [80,201,202,373], matrix product and tensor network states [374–376], and techniques that assume a gate's errors can be described by few-parameter (i.e., sparse) Pauli channels [377] or Lindbladians [378,379]. Methods also exist that use heuristics to improve efficiency, such as some machine learning approaches to tomography [380–384]. With the exception of some efficient Pauli-noise learning methods covered in Sec. IX C, we do not discuss these techniques further.

A. Truth table tomography

Truth table tomography [385,386] is perhaps the conceptually simplest form of partial tomography of an n-qubit process Λ. It consists of preparing the n qubits in each of the 2^n different computational basis states, applying Λ, and then measuring in the computational basis. This directly estimates the 4^n different probabilities given by

$$p_{y|x}(\Lambda) = \mathrm{Tr}[|y\rangle\langle y|\,\Lambda(|x\rangle\langle x|)]\,, \tag{391}$$

where $x, y \in \{0,1\}^n$. The matrix of these probabilities $p_{y|x}(\Lambda)$ is a $2^n \times 2^n$ *stochastic matrix* (each element of the matrix is a probability, and each row sums to one), which we denote by $S(\Lambda)$. This matrix is similar to the response (or confusion) matrix constructed when characterizing readout fidelities [see Eq. (318)].

As in quantum process tomography (QPT, see Sec. VII B), the measured stochastic matrix $S(\Lambda)$ is typically compared to the stochastic matrix $S(\mathcal{U})$ for the intended (ideal) superoperator $\mathcal{U}$ of a gate. For any gate that preserves the computational basis (i.e., each computational basis state is mapped to another computational basis state), $S(\mathcal{U})$ is a *truth table*, i.e., it is the matrix for a deterministic (reversible) classical gate. For example, for the CNOT gate this matrix is

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{392}$$

Note, however, that the stochastic matrix for a general unitary is instead a general (doubly [387]) stochastic matrix, e.g., for a Hadamard gate

$$S = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{393}$$

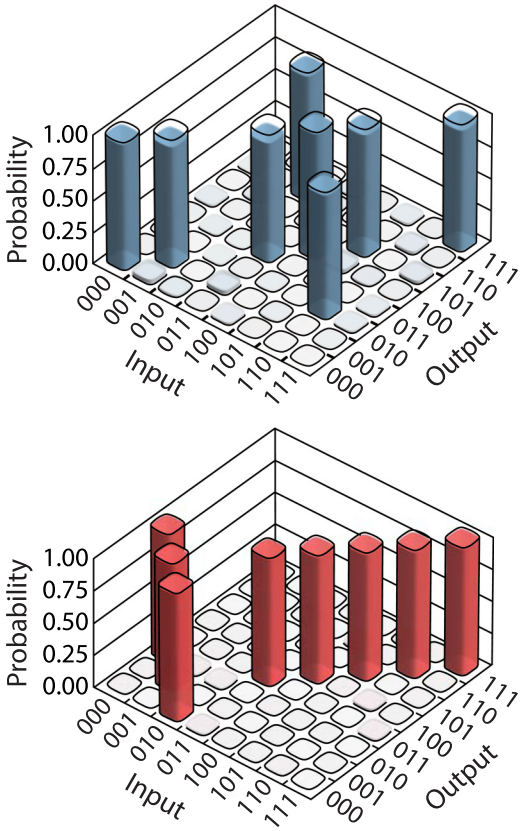


FIG. 37. Truth table tomography. The stochastic matrices for experimental (top) Toffoli [218] and (bottom) *i*Toffoli gates [263] measured using truth table tomography on superconducting qubits. The Toffoli gate flips the first qubit if the second and third qubits are in the state $|10\rangle$. The *i*Toffoli gate flips the middle qubit if the first and third qubit are in the state $|00\rangle$. The corresponding gate fidelities are estimated to be 96.20(6)% and 98.6(1)%, respectively.

To demonstrate truth table tomography, in Fig. 37 we plot measured and ideal stochastic matrices for experimental Toffoli [218] and *i*Toffoli [263] gates. Ideal Toffoli and *i*Toffoli gates preserve the computational basis. The Toffoli gate leaves the qubits unchanged unless the two control qubits (in this experiment, the second two qubits) are in the state $|10\rangle$, in which case the other qubit (in this experiment, the first qubit) is flipped, i.e., $|010\rangle \mapsto |110\rangle$ and $|110\rangle \mapsto |010\rangle$. The *i*Toffoli gate leaves the qubits unchanged unless the two control qubits (in this experiment, the first and last qubit) are in the state $|00\rangle$, in which case the other qubit (in this experiment, the middle qubit) is flipped, i.e., $|000\rangle \mapsto |010\rangle$ and $|010\rangle \mapsto |000\rangle$.

For any gate that ideally maps computational basis states to computational basis states, the fidelity between an experimental and ideal gate's stochastic matrices, S_{exp} and S_{ideal}, respectively, is given by [385,386]

$$F_{\mathrm{tt}} = \frac{1}{2^n}\mathrm{Tr}(S_{\mathrm{exp}}^{\mathrm{T}} S_{\mathrm{ideal}})\,. \tag{394}$$

For the experimental gates of Fig. 37, we find fidelities of 96.20(6)% and 98.6(1)% for the Toffoli and *i*Toffoli gates, respectively.

Truth table tomography has a variety of limitations. Like full QPT, it requires a number of circuits that scales exponentially in the number of qubits. Furthermore, like QPT it is susceptible to SPAM errors. However, unlike QPT, it is insensitive to phase errors in the gates. Therefore, fidelities measured using truth table tomography will typically disagree with those gate fidelities estimated using other techniques, such as interleaved RB (Sec. VIII F) or cycle benchmarking (Sec. VIII G) [218]. To recover some information about the phases of a gate, the input states can be rotated to the X basis. In combination with the Z-basis results, this data can be used to lower bound the process fidelity [388], and this has been used to characterize the effects of three-qubit Toffoli gates in trapped ions [389] and neutral atoms [390]. Moreover, the method introduced in Ref. [388] can be generalized to upper- and lower-bound the fidelity of high-dimensional operations such as n-qubit Toffoli gates [391]. However, this is an inefficient approach to estimating gate fidelity compared to, e.g., direct fidelity estimation.

B. Direct fidelity estimation

Estimating the fidelity of a state or process requires computing the overlap between a system's state and process with the desired state and process (see Secs. IV B 2 and IV C 3 b). While, in principle, a full tomographic reconstruction of the system could be used to accurately compute its overlap with the desired output, in practice full tomography becomes intractable beyond a few qubits. However, this overlap can be estimated by measuring only along axes of greater overlap with the desired state, while neglecting axes with little to no overlap. *Direct fidelity estimation (DFE)* [367,368] is a protocol that uses this idea to measure state or process fidelities more efficiently than full tomography [392,393]. While DFE can in theory be used to compute the fidelity of entire circuits, more scalable methods have been developed specifically for estimating circuit fidelities (see Sec. X).

The purpose of DFE is to estimate the state fidelity [Eq. (198)] between an actual state ρ and a desired pure state $\psi = |\psi\rangle\langle\psi|$. This fidelity can be written as

$$F(\rho,\psi) = \mathrm{Tr}[\psi\rho] = \sum_{k=1}^{d^2} \chi_\psi(k)\chi_\rho(k)\,, \tag{395}$$

where

$$\chi_\rho(k) = \frac{\mathrm{Tr}[\rho P_k]}{\sqrt{d}} \tag{396}$$

is known as the *characteristic function* of ρ, P_k are the n-qubit Pauli operators, and $d = 2^n$ (for n qubits). The quantity $\chi_\rho(k)$ is the k^{th} expansion coefficient of ρ in the normalized Pauli basis. While the exact expansion of Eq. (396) includes d^2 terms, $F(\rho,\psi)$ can be estimated by measuring only a subset of the most significant terms using importance sampling. To estimate the fidelity up to an additive error ϵ and failure probability δ, one can take the following steps:

(a) Choose a random value $k \in \{1,\ldots,d^2\}$ with probability $p_k = \chi_\psi(k)^2$.
(b) Calculate $\chi_\rho(k)$ by measuring the expectation value of the Pauli operator P_k for the unknown state ρ. Use this quantity to construct the estimator $X = \chi_\rho(k)/\chi_\psi(k)$.
(c) Repeat the steps above $l = 1/(\epsilon^2\delta)$ times and estimate $F(\rho,\psi)$ by the estimator

$$\hat{F}(\rho,\psi) = \frac{1}{l}\sum_{i=1}^{l} X_i\,. \tag{397}$$

Reference [367] proves that if $\chi_\rho(k)$ is measured exactly, then

$$\Pr[|\hat{F}(\rho,\psi) - F(\rho,\psi)| \geq \epsilon] \leq \delta\,. \tag{398}$$

However, there is always shot noise, i.e., each $\chi_\rho(k)$ is not measured perfectly. Reference [367] shows that ρ's fidelity can be estimated using m_k copies of ρ for each P_k. While the exact number of copies varies with k, the average number of copies grows only linearly in d (as opposed to quadratically, as for full state tomography). Specifically, if we have $m = \sum_{i=1}^{l} m_i$ total copies of ρ, then the expected number of copies required for a given δ and ϵ is

$$\mathbb{E}(m) \leq 1 + \frac{1}{\epsilon^2\delta} + \frac{2d}{\epsilon^2}\log(2/\delta)\,. \tag{399}$$

Thus, we can generally reduce the cost from $d^2 \longrightarrow d$ by using DFE instead of state tomography.

The average number of copies can be significantly decreased for particular families of states. For example, let us consider the family of "well-conditioned" states, which includes all the states ρ such that for every k, either $\mathrm{Tr}[\rho P_k] = 0$ or $|\mathrm{Tr}[\rho P_k]| \geq \alpha$ for $\alpha \leq 1$. For states in this family, we have

$$m \leq \mathcal{O}\left(\frac{\log(1/\delta)}{\alpha^2\epsilon^2}\right). \tag{400}$$

For example, for stabilizer states $\alpha = 1$, so the number of copies of ρ needed is independent of the system's size, and for W states ($\alpha = 1/n$), the average number of copies grows only as n^2.

DFE is cheaper than tomography, but its cost is still exponential in the number of qubits (n) for general states. Furthermore, like state tomography, DFE does not account

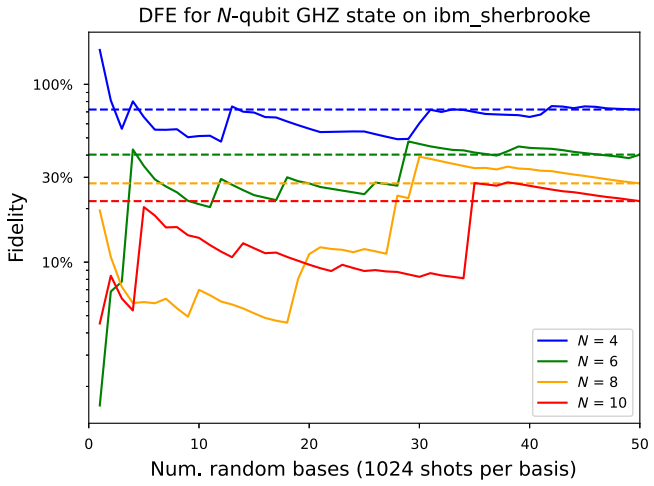


FIG. 38. Direct fidelity estimation. DFE performed on an $N = 4, 6, 8, 10$ qubit GHZ state on `ibm_sherbrooke`. The running average of the GHZ state fidelity is plotted as a function of the number of random measurement bases, selected via importance sampling on the target state. The horizontal dashed lines is the expected state fidelity calculated from the physical error rates listed for `ibm_sherbrooke` at the time of the experiment.

for measurement errors, so DFE's estimates of state fidelity will also include errors from measurement. For this reason, fidelity estimation techniques which account for SPAM have been developed (see Sec. X A).

To demonstrate how DFE depends on the number of random bases that are sampled, Fig. 38 shows results from DFE of an n-qubit Greenberger-Horne-Zeilinger (GHZ) state (for $n = \{4, 6, 8, 10\}$) performed on `ibm_sherbrooke`. Rather than measuring 1 shot for each sampled basis, as described above, the DFE prescription from Ref. [364] was followed. 1024 shots from each of 50 measurement bases randomly drawn via importance sampling from the ideal target GHZ state were measured. The running average of the fidelity for each GHZ state is shown in Fig. 38. The estimated fidelities after 50 random measurement bases are in general agreement with the expected output fidelity based on a product-of-errors calculation for the physical error rates on `ibm_sherbrooke`.

C. Pauli noise learning

Stochastic Pauli channels (see Sec. III E) are an important class of error channels that are particularly relevant for quantum error correction [394]. A general Pauli error map for n qubits only contains $4^n - 1$ parameters—the rates of each possible Pauli error—which is fewer than the $\mathcal{O}(16^n)$ parameters of a general process matrix. While this is still exponential in the number of qubits, it can be further reduced using assumptions about the nature and locality of Pauli errors across an n-qubit device. Furthermore, while most quantum systems suffer from more complex error mechanisms than simply Pauli noise, one can experimentally *design* stochastic channels [395] using methods such as randomized compiling [154,155] and Pauli frame randomization [152,153], thus enforcing the same error model that can be efficiently characterized.

Several methods have been proposed for learning Pauli channels [335,396]. Pauli channels have diagonal Pauli transfer matrices (PTMs, see Sec. II C 3), denoted Λ, and these methods are designed to estimate the eigenvalues

$$\Lambda_{PP} = \frac{1}{d}\mathrm{Tr}[P\mathcal{E}(P)]\,. \tag{401}$$

The eigenvalue $\lambda_P = \Lambda_{PP}$ captures how much $\mathcal{E}$ attenuates the Pauli operator P. If $\Lambda_{PP} = 1$, then $\mathcal{E}$ preserves the Pauli operator P; if $\Lambda_{PP} < 1$, then P is not preserved by $\mathcal{E}$.

The eigenvalues Λ_{PP} can be related to the rates of each possible Pauli error, as we demonstrate using a one-qubit Pauli channel, which has the form

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2(p_Y+p_Z) & 0 & 0 \\ 0 & 0 & 1-2(p_X+p_Z) & 0 \\ 0 & 0 & 0 & 1-2(p_X+p_Y) \end{pmatrix}, \tag{402}$$

where p_Q is the probability of the Pauli error Q. This example shows that a Pauli error Q with probability p_Q will attenuate the eigenvalue of any noncommuting Pauli operator P by an amount $2p_Q$. To generalize the relationship between Pauli eigenvalues and Pauli error rates, we note that a stochastic Pauli channel's Kraus map is of the form $\mathcal{E}(\rho) = \sum_Q p_Q Q\rho Q^\dagger$, and this has a PTM given by

$$\Lambda = \frac{1}{d}\sum_{P,Q\in\mathbb{P}_n} (-1)^{\langle P,Q\rangle} p_Q \,|P\rangle\!\rangle\langle\!\langle P|\,, \tag{403}$$

where $\langle P, Q\rangle = 0$ if $[P, Q] = 0$, otherwise $\langle P, Q\rangle = 1$, and $|\cdot\rangle\!\rangle$ is the *vectorization* defined in Sec. II C. A single given eigenvalue Λ_{PP} may therefore be computed as

$$\Lambda_{PP} = \sum_{Q\in\mathbb{P}_n} (-1)^{\langle P,Q\rangle} p_Q\,. \tag{404}$$

The inverse transformation—computing a Pauli error rate p_Q from Pauli eigenvalues—is

$$p_Q = \frac{1}{4^n}\sum_{P\in\mathbb{P}_n} (-1)^{\langle P,Q\rangle}\Lambda_{PP}\,. \tag{405}$$

This transformation is the *Walsh-Hadamard* transform, with the following matrix representation:

$$\mathbb{W}_{P,Q} = \frac{1}{d}\sum_{P,Q}(-1)^{\langle P,Q\rangle}\,|P\rangle\!\rangle\langle\!\langle Q|\,. \tag{406}$$

This Walsh-Hadamard transform can be used to compute a vector of Pauli eigenvalues $\boldsymbol{\lambda}$ from a vector of Pauli error rates $\mathbf{p}$,

$$\boldsymbol{\lambda} = \mathbb{W}\mathbf{p}\,, \tag{407}$$

or, using the inverse transformation, to compute Pauli error rates from Pauli eigenvalues:

$$\mathbf{p} = \mathbb{W}^{-1}\boldsymbol{\lambda}\,. \tag{408}$$

Thus, the general strategy for learning Pauli channels is to measure a set of Pauli eigenvalues and then use Eq. (409) to calculate the associated Pauli error rates. Measuring Pauli eigenvalues can be achieved using various methods, including cycle benchmarking (Sec. VIII G), Pauli-twirled random Clifford circuits (see Sec. IX C 2), shadow tomography [397], etc. In what follows, we will describe two strategies for estimating Pauli eigenvalues.

1. Cycle error reconstruction

In Sec. VIII G, we reviewed cycle benchmarking (CB), a scalable protocol for measuring error rates for cycles containing parallel quantum gates. The goal of CB is to measure the eigenvalues of the PTM of a cycle. These eigenvalues can be estimated from a Pauli-twirled PTM; for this reason, CB uses randomized compiling. For example, CB performed on a two-qubit cycle would measure the following diagonal components of the PTM (up to any gauge ambiguities, see Appendix E),

$$\Lambda = \begin{pmatrix} 1 & & & & & & & & & & & & & & & \\ & f_{IX}^m & & & & & & & & & & & & & & \\ & & f_{IY}^m & & & & & & & & & & & & & \\ & & & f_{IZ}^m & & & & & & & & & & & & \\ & & & & f_{XI}^m & & & & & & & & & & & \\ & & & & & f_{XX}^m & & & & & & & & & & \\ & & & & & & f_{XY}^m & & & & & & & & & \\ & & & & & & & f_{XZ}^m & & & & & & & & \\ & & & & & & & & f_{YI}^m & & & & & & & \\ & & & & & & & & & f_{YX}^m & & & & & & \\ & & & & & & & & & & f_{YY}^m & & & & & \\ & & & & & & & & & & & f_{YZ}^m & & & & \\ & & & & & & & & & & & & f_{ZI}^m & & & \\ & & & & & & & & & & & & & f_{ZX}^m & & \\ & & & & & & & & & & & & & & f_{ZY}^m & \\ & & & & & & & & & & & & & & & f_{ZZ}^m \end{pmatrix}, \tag{409}$$

where f_P^m is the process polarization associated with the preparation and measurement basis of the Pauli P at a circuit depth of m. Here, f_P is distinct from the process polarization f defined in Sec. IV C 4, which is the *average* process polarization in the unital block of a PTM. Under Clifford twirling, $f = 1/(d^2-1)\sum_{P\neq I} f_P$, but under Pauli twirling each f_P is unique (see Sec. VIII A and Appendix C). In the limit that $m = 0$, f_P is an estimate of eigenvalue Λ_{PP}, representing how well P is preserved by $\mathcal{E}$. It should be noted that f_P is often termed the *Pauli fidelity* of P in the literature when P is unambiguous [398], or the *orbital fidelity* when individual Pauli fidelities cannot be learned due to gauge ambiguities; however, f_P is not strictly a fidelity because it can be negative, since the elements of a PTM are bounded by $[-1, 1]$.

Cycle error reconstruction (CER) [396,398] (also called *k-body noise reconstruction (KNR)* [343]) is a protocol that leverages CB for efficiently estimating the eigenvalues of a cycle's PTM. CER results are based on targeted CB measurements in which specific Paulis are chosen to estimate the error rates afflicting subsets of the gates or idle qubits in the specific cycle of interest. Since the Pauli decays in CB are dual to the Pauli operators that cause errors, CER measures the error rate p_Q of some fixed Pauli Q, by measuring a set of Pauli decays that commute and anticommute with Q, and then using this information to reconstruct the probability p_Q via linear inversion using Eq. (409).

For the two-qubit PTM shown in Eq. (410), it is feasible to measure all 15 operators, and thus reconstruct all weight-1 and weight-2 Pauli errors afflicting the cycle.

However, for an arbitrary n-qubit cycle, while some high-weight errors can be estimated, it becomes exponentially expensive to measure all $4^n - 1$ Pauli errors. Instead, the usual strategy is to only reconstruct lower-weight Pauli errors, thus limiting the number of Pauli eigenvalues that must be measured. This strategy assumes that errors are relatively local to nearby qubits, and that long-range correlations are negligibly small.

In Fig. 39, we plot a heatmap of the dominant Pauli errors on an eight-qubit superconducting quantum processor with a ring geometry. These errors are reconstructed by performing CB on the eight different cycles containing a single two-qubit CZ gate, as well as the idle qubits on either side of each CZ (i.e., the interleaved gate cycle is $G = I \otimes \mathrm{CZ} \otimes I$, see Fig. 34). We see that the dominant Pauli errors on the quantum processor are weight-1 errors affecting the idle qubits or one of the entangled qubits. In fact, the largest error on the processor is a local Z error on qubit 4 during the CZ gate between qubits 3 and 4. The source of this error is likely to a *coherent* Z error, not a *stochastic* Z error (by design, CER cannot distinguish between true stochastic Pauli errors and coherent errors that have been twirled into Pauli channels).

In Fig. 39, we observe that some of the Pauli errors acting on entangled qubits appear grouped together in curly brackets. These groupings indicate error types that cannot be distinguished due to degeneracies, since some local errors acting on either qubit in the CZ gate will be transformed by the gate. This has to do with a fundamental gauge ambiguity in Pauli noise learning [399], and is explained further in Appendix E. While only one- and two-body errors [400] were measured in Fig. 39 (i.e., all $k \geq 2$-body errors were neglected), this is justified by the data, since we observe that two-body terms are largely suppressed compared to one-body terms. Moreover, the two-body error rates are the marginalized probabilities of *all* k-body errors that act on the corresponding two bodies. Therefore, the fact that two-body errors are negligible proves that three- or more body errors are also negligible.

2. Averaged circuit eigenvalue sampling

Averaged circuit eigenvalue sampling (ACES) [377, 401,402] is an alternative scalable technique for learning the Pauli error rates of many layers of gates performed simultaneously. ACES uses *random* Clifford circuits performed with randomized compiling to accomplish this, which allows many gates to be characterized in a single experiment.

ACES estimates a Pauli channel for each gate in a set of Clifford gates. ACES can estimate arbitrary n-qubit Pauli error rates for each gate in principle, but, like CER, a reduced model is required for scalability. For example, a crosstalk-free error model [78,79] can be used, in which each layer's error consists of tensor products of one- and two-qubit Pauli channels for each one- and two-qubit gate in the layer, respectively. Even for this highly restricted error model, the parameter space becomes large very quickly—this model has $15N_{2Q} + 3N_{1Q}$ parameters, where N_{2Q} is the number of 2-qubit gates and N_{1Q} is the number of single-qubit gates in the layer. For example, for a line of 100 qubits with bidirectional CNOT gates and six single-qubit Clifford gates per qubit, this amounts to 4770 parameters. It is also possible to incorporate additional variables for state preparation and measurement error.

ACES uses measurements of Pauli observables of random Clifford circuits to learn many combinations of the eigenvalues of Pauli channels, which can then be used to estimate the individual eigenvalues themselves. Consider a circuit $\mathcal{C} = G_d G_{d-1} \cdots G_1$, where the G_i are Clifford gates, and suppose each gate experiences a (gate-dependent) post gate stochastic Pauli error $\Lambda_{\mathcal{E}}^{G_i}$, i.e., the noisy circuit is

$$\tilde{\mathcal{C}} = \Lambda_{\mathcal{E}}^{G_d} \Lambda_d \cdots \Lambda_{\mathcal{E}}^{G_1} \Lambda_1 , \tag{410}$$

where Λ_i denotes the PTM for G_i and $\Lambda_{\mathcal{E}}^{G}$ denotes the superoperator of G's error channel. A Pauli measurement result is determined by the *generalized eigenvalues* of the imperfect gates,

$$\Lambda_{\mathcal{E}}^{G} \Lambda[P] = \left(\Lambda_{\mathcal{E}}^{G}\right)_{GPG^{-1}, GPG^{-1}} \left[GPG^{-1}\right] . \tag{411}$$

Stated differently, each Clifford operation transforms a Pauli P into another Pauli P' (see Appendix B 3), and P' is always an eigenvector of the subsequent Pauli channel. By applying Eq. (412) to a sequence of gates, we see that the Pauli operators are generalized eigenvectors of any Clifford circuit that experiences only stochastic Pauli noise, and the circuit's generalized eigenvalues, denoted $\lambda_P^{\mathcal{C}}$, are

$$\Lambda_{\mathcal{E}}^{G_d} \Lambda_d \cdots \Lambda_{\mathcal{E}}^{G_1} \Lambda_1[P] = \prod_{i=1,\ldots,m} \left(\Lambda_{\mathcal{E}}^{G_i}\right)_{P_i, P_i} \left[\mathcal{C}P\mathcal{C}^{-1}\right] , \tag{412}$$

$$= \lambda_P^{\mathcal{C}} \left[\mathcal{C}P\mathcal{C}^{-1}\right] \tag{413}$$

where P_i denotes the Pauli P evolved through the first i gates, i.e., $P_i = \Lambda_i \cdots \Lambda_1[P]$. By (1) preparing properly sampled eigenstates of P (see Sec. IX B), (2) performing $\mathcal{C}$, and then (3) measuring the final Pauli $\mathcal{C}P\mathcal{C}^{-1}$, the generalized eigenvalues $\lambda_P^{\mathcal{C}}$ can be determined experimentally.

To use the Pauli measurement results to estimate individual gate eigenvalues λ_P^G (and hence the error model parameters), we construct a linear system of equations relating the circuit generalized eigenvalues to the λ_P^G. Taking the log of Eq. (413) (and assuming $\lambda_{P_i}^{G_i} > 0$ for all i, which is true as long as the gates have sufficiently low error

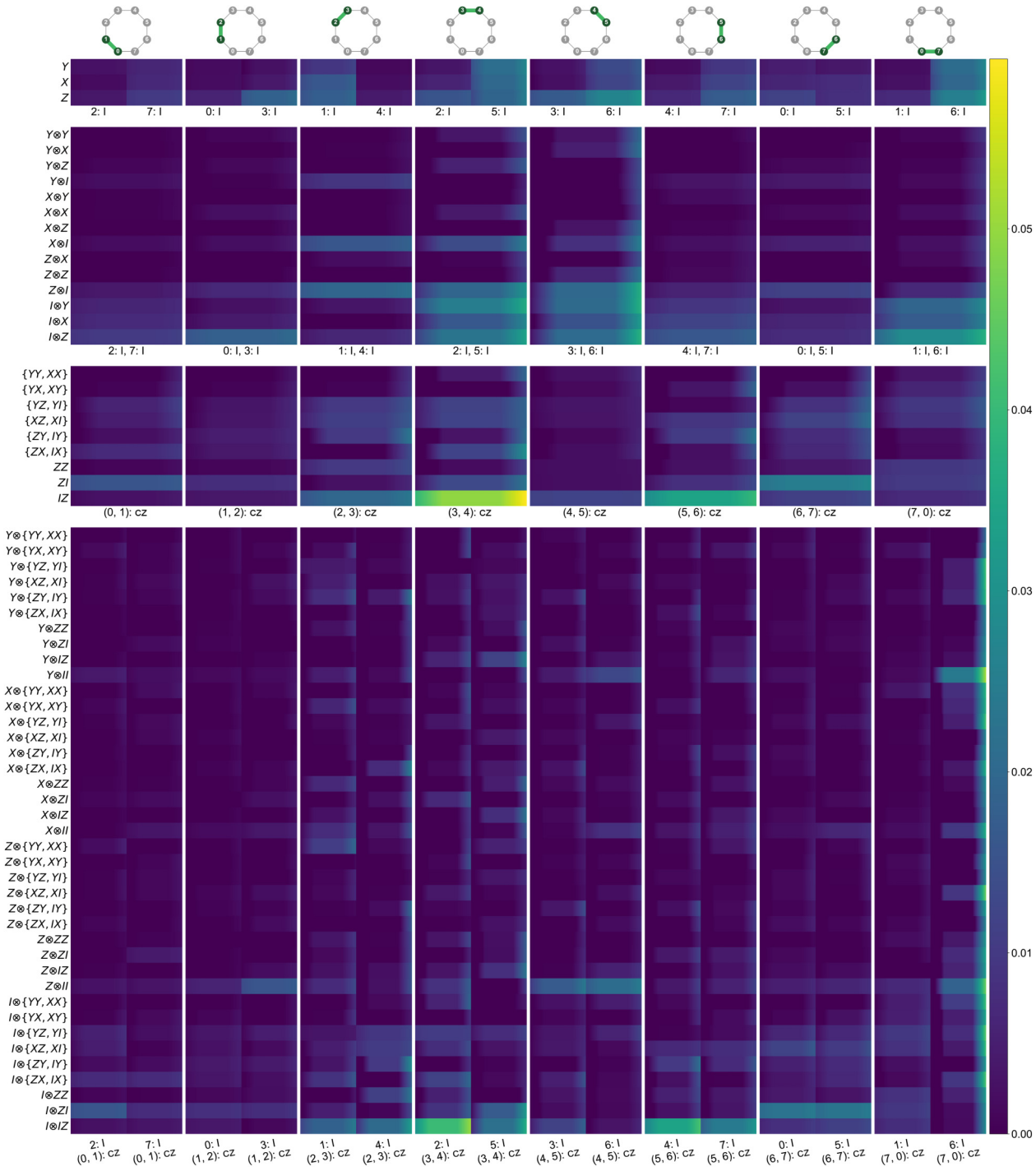


FIG. 39. Cycle error reconstruction. Heatmap of Pauli errors measured via CER for an eight-qubit superconducting quantum processor. Each experiment consisted of benchmarking a two-qubit CZ gate (indicated by the graph pattern above each column) plus nearest-neighbor idling qubits. The x axis labels the ideal gate operation on each subset of qubits in each benchmarked cycle. The y axis labels the type of Pauli error, with the tensor notation $\otimes$ between Pauli errors indicating errors acting on product states, while the lack of tensors indicates errors on entangled qubits; curly brackets indicate gauge ambiguities (see Appendix E). The color of each cell indicates the marginalized error rate, and the gradient defines the 95% confidence interval. The first row of subplots shows single-body errors acting on idling qubits; the second row of subplots shows correlated single- and two-body errors between idling qubits; the third row of subplots shows single-body errors acting on the CZ qubits; and the fourth row of subplots shows correlated single- and two-body errors between an idling spectator qubit and the CZ gate qubits.

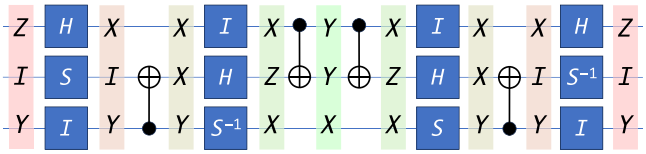


FIG. 40. Averaged circuit eigenvalue sampling (ACES). Example mirror circuit that can be used for the ACES protocol, consisting of single-qubit Clifford gates plus bidirectional CX gates. The three qubits are prepared in a $+1$ eigenstate of *ZIY*. Each layer of gates transforms the Pauli operator *ZIY* into a different Pauli, and this Pauli describes which of the layer's generalized eigenvalues will contribute to the final measurement result. Under noise-free execution, the final state is the same as the input state, and therefore is a $+1$ eigenstate of *ZIY*. With execution under a stochastic Pauli noise model, the result of measuring *ZIY* is attenuated by $\lambda^{\mathcal{C}}_{ZIY}$, which is the circuit eigenvalue.

rates),

$$\ln \lambda_P^C = \sum_{j=1}^{d} \ln \left[\left(\Lambda_{\mathcal{E}}^{G_j} \right)_{P_j, P_j} \right]. \tag{414}$$

Therefore, Pauli measurement results are related to the generalized eigenvalues of the gates by a system of linear equations $\mathbf{b} = A\mathbf{x}$, where $\mathbf{x}$ encodes the Pauli channel eigenvalues and $\mathbf{b}$ encodes the measurement results. A is called the *design matrix*, and each row of A encodes how the gate eigenvalues relate to the result of a single Pauli observable measurement. It is determined by the choice of circuits and Pauli measurements, and can be efficiently computed since its computation only requires evolving Pauli operators through Clifford circuits. By running sufficiently many random Clifford circuits and performing sufficiently many independent Pauli measurements, a full-rank matrix A can be generated. The Pauli eigenvalues of the gates can then be estimated by computing $\mathbf{x} = A^{+}\mathbf{b}$, where A^{+} denotes the pseudoinverse of A.

In principle, ACES can be run with any set of Clifford circuits, but because most processors are limited to computational basis measurements, a careful choice of circuit structure allows for more independent Pauli measurements to be performed with each circuit, rendering more information. ACES is often run with a form of mirror circuit (see Fig. 40), which enables measuring $\mathcal{O}(n)$ independent Pauli observables from each computational basis measurement. The version of mirror circuits in the original ACES protocol (Ref. [377]) places a layer of random gates at the end of the circuit, so they are not identity circuits (or Pauli operators). The particular structure used means that any one- or two-qubit Pauli propagated through the full circuit has weight at most 6 at the end of the circuit, which means that each measurement required to learn a crosstalk-free model with ACES requires at most six qubits.

X. ESTIMATING CIRCUIT FIDELITIES

While all the individual components (states, gates, layers, and measurements) used in quantum circuits are becoming more accurate, they remain inherently noisy. When a few of them are combined to form a quantum circuit, their noise and errors accumulate. And when more than a few of them are combined (hundreds, thousands, or even—eventually—millions), the accumulated noise becomes significant and can severely alter the outputs. So, given a quantum circuit of arbitrary size and nature, how can we determine whether its outputs are close to the ideal (noiseless) outputs?

Validating the outputs of quantum circuits turns out to be a puzzling problem. The simplest approach is to resort to classical simulations: when a quantum circuit is implemented, it is also simulated on a classical computer, and finally the outputs are compared. This approach is effective, but only for circuits that are feasible to simulate classically. Some experiments have already hit the boundary of what can be simulated classically in reasonable time [2]. A different approach consists of individually characterizing the components used in the circuit of interest (e.g., benchmarking of gate layers, or tomography of individual gates), and then predicting the quality of its outputs from the characterization data. This approach can be scalable, but it is also often unreliable. Quantum circuits are more than the sum of their components, and the noise in a circuit may exhibit properties (such as drift, fluctuations, and temporal correlations) that may not be observed by inspecting individual components. This calls for protocols that can test the circuit as a whole, rather than its parts.

In this section, we provide an overview of some of the scalable methods to characterize the performance of quantum circuits. In particular, we describe the following methods:

(a) *Mirror circuit fidelity estimation* (Sec. X A). Mirror circuit fidelity estimation is a technique for estimating the process fidelity of any circuit $\mathcal{C}$ using mirror circuits of twice $\mathcal{C}$'s depth.
(b) *Circuit output accreditation* (Sec. X B). Circuit output accreditation is a technique for lower bounding the process fidelity of a circuit $\mathcal{C}$ by running a set of "trap" circuits that are the same width and depth as $\mathcal{C}$, but contain only Clifford gates.

Mirror circuit fidelity estimation and circuit output accreditation are complementary techniques with similar aims and slightly different properties and strengths (discussed later). These techniques have two important properties in common: they are both (i) efficient in the number of qubits, and (ii) robust to SPAM errors. Both techniques run a number of circuits that is independent of the number of qubits, and require minimal classical computations. This

contrasts with direct fidelity estimation (Sec. IX B), which could be used to estimate a circuit's process fidelity, but is typically not used in practice because it is not robust to SPAM errors and it is exponentially expensive for general circuits. Finally, note that a variety of techniques exist for formal verification of the output of quantum algorithms or circuits, for example, interactive cryptographic protocols, which allow a classical user to verify that a computation was carried out by a quantum device [403–406]. These methods are beyond the scope of this tutorial.

A. Mirror circuit fidelity estimation

Mirror circuit fidelity estimation (MCFE) [318] is a technique for efficiently measuring the process (i.e., entanglement) fidelity $F_e(\mathcal{C})$ with which a quantum computer can implement an n-qubit circuit $\mathcal{C}$. MCFE is robust in the presence of SPAM errors, and it is efficient in the number of qubits. MCFE consists of running circuits sampled from three ensembles of "mirror circuits" built from the circuit of interest $\mathcal{C}$, shown in Fig. 41. The core idea is that by running circuits with three different structures, one of which contains $\mathcal{C}$, MCFE is able to approximately isolate the process fidelity of $\mathcal{C}$ from all other operations in those circuits using some simple algebra. Below we explain how MCFE works.

MCFE's first mirror circuit ensemble $[M_1(\mathcal{C})]$ consists of (i) a layer L of random single-qubit gates sampled from a unitary 2-design (e.g., Haar-random single-qubit gates), (ii) the circuit $\mathcal{C}$, (iii) a randomly compiled version of the inverse of $\mathcal{C}$ [154,155] (denoted $\mathcal{C}_{\text{rev}}$), and (iv) the inverse of L compiled together with a random n-qubit Pauli gate. Each such $M_1(\mathcal{C})$ circuit embeds $\mathcal{C}$ within a larger (mirror) circuit that, if implemented without error, will always return an easy-to-compute "success" bit string. Therefore, we can easily assess how well each such circuit was run simply by looking at the frequency with which this success bit string is output from each $M_1(\mathcal{C})$ circuit—suggesting that these circuits can be used to understand how well $\mathcal{C}$ can be executed. In particular, because of the randomization in the initial and final layer of gates, as well as the randomized compilation in $\mathcal{C}_{\text{rev}}$, it is possible to show that

$$\mathbb{E}(\gamma[M_1(\mathcal{C})]) \approx f_0 f(\mathcal{C}_{\text{rev}}) f(\mathcal{C}) . \tag{415}$$

where $f(\cdot)$ is the process polarization [see Eq. (245)], $\mathbb{E}(\cdot)$ denotes the expectation value over a circuit ensemble,

$$\gamma(M) = \frac{4^n}{4^n - 1} \sum_{k=0}^{n} \left(-\frac{1}{2}\right)^k h_k(M) - \frac{1}{4^n - 1} , \tag{416}$$

where $h_k(M)$ is the frequency with which the output of mirror circuit M is a Hamming distance of k from its "success" bit string, and f_0 is a nuisance parameter called the "effective SPAM polarization," which encompasses contributions from errors in the SPAM and in the layers of single-qubit gates [steps (i) and (iv)].

The aim in MCFE is to measure $F_e(\mathcal{C})$ [a rescaling of $f(C)$, see Table II], but if we only run circuits sampled from $M_1(\mathcal{C})$ to estimate $\mathbb{E}(\gamma[M_1(\mathcal{C})])$, we instead learn $f(\mathcal{C})$ multiplied by the unknowns f_0 and $f(\mathcal{C}_{\text{rev}})$. MCFE solves this problem by running circuits sampled from two additional ensembles $[M_2(\mathcal{C})$ and $M_3(\mathcal{C})]$. $M_3(\mathcal{C})$ is essentially a randomized SPAM experiment [steps (i) and (iv) above] that enables learning f_0:

$$\mathbb{E}(\gamma[M_3(\mathcal{C})]) = f_0 . \tag{417}$$

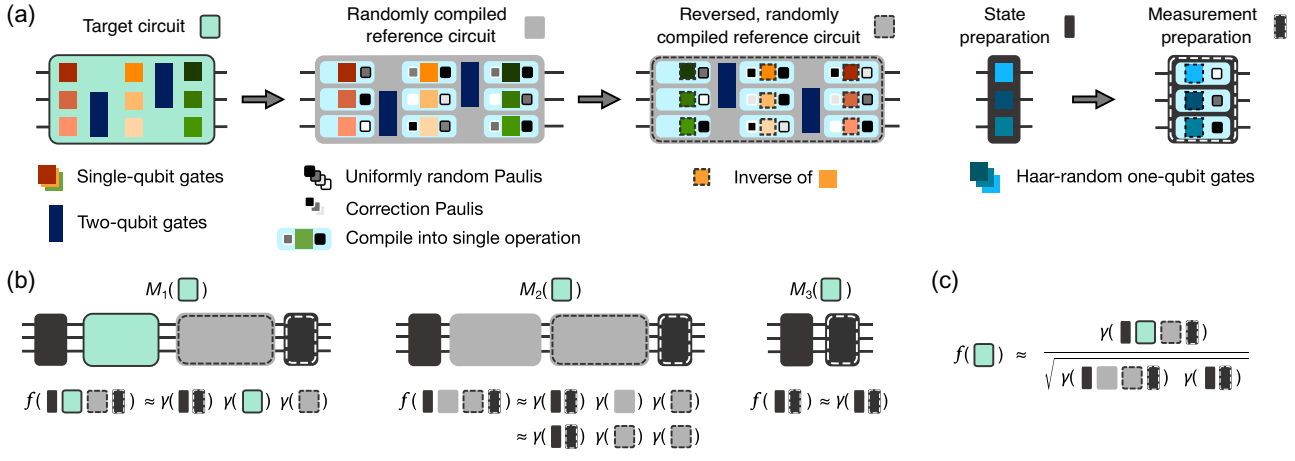


FIG. 41. Mirror circuit fidelity estimation. (a) MCFE estimates the process fidelity $F_e(\mathcal{C})$ with which a quantum computer can execute some target circuit $\mathcal{C}$ (green box), using motion reversal circuits built from $\mathcal{C}$ and the four reference (sub)circuits shown here. (b) The three motion reversal circuits used in MCFE and (c) the simple data analysis that MCFE uses to estimate the $\mathcal{C}$'s process polarization $f(\mathcal{C})$, which can be rescaled to estimate $F_e(C)$ (see Table II) (Figure adapted with permission from Ref. [318]).

Finally, $M_2(\mathcal{C})$ is a fully randomly compiled version of $M_1(\mathcal{C})$ (see Fig. 41), which enables learning $f(\mathcal{C}_{\text{rev}})$:

$$\mathbb{E}(\gamma[M_2(\mathcal{C})]) \approx f_0 f(\mathcal{C}_{\text{rev}})^2 . \tag{418}$$

By applying simple algebra to Eqs. (416), (418), and (419), we see that

$$f(\mathcal{C}) \approx \frac{\mathbb{E}(\gamma[M_1(\mathcal{C})])}{\sqrt{\mathbb{E}(\gamma[M_2(\mathcal{C})])\mathbb{E}(\gamma[M_3(\mathcal{C})])}} . \tag{419}$$

This is the analysis used by MCFE to estimate $f(\mathcal{C})$, which can then be rescaled to estimate $F_e(\mathcal{C})$ using Eq. (245).

B. Circuit output accreditation

Circuit output accreditation is an efficient strategy for lower-bounding the process fidelity in a "target" circuit of interest. It only requires implementing circuits with the same size and depth as the target circuit. Moreover, it is robust to SPAM errors, and it is scalable in the number of qubits and gates in the target circuit.

Different variants of circuit output accreditation have been proposed [407–409], but they all rely on the idea of implementing the target circuit alongside a number N_c of Clifford circuits, called "traps" (see Fig. 42). These traps have the same width and depth as the target circuit, but they implement different computations. In particular, the traps are designed in such a way that, in the absence of noise, they return a fixed, known output. This allows us, in the presence of noise, to estimate the probability that a trap returns an incorrect output. This probability can then be used to bound the process fidelity of the target circuit. To describe circuit accreditation in more detail, we focus on the protocol in Ref. [409], which provides the tightest bound on the fidelity of the target circuit.

The accreditation protocol takes as input a target circuit $\mathcal{C}$, alongside two numbers $\theta, \alpha \in (0, 1)$, which represent the desired statistical error on the bound and the confidence in the bound, respectively. To bound the process fidelity $F_e(\mathcal{C})$ of $\mathcal{C}$, output accreditation makes the following assumptions:

(1) The circuit $\mathcal{C}$ (i) takes as input n qubits in the state $|0\rangle$, (ii) implements the sequence of operations $U_{m+1}E_mU_m \ldots U_2E_1U_1$, where U_j is a layer of single-qubit gates and E_j is a layer of two-qubit (e.g., CZ) gates for every j, and (iii) ends with Pauli-Z measurements on every qubit.
(2) The errors affecting the various layers in $\mathcal{C}$ are completely positive and trace-preserving (CPTP).
(3) The errors affecting the layers of one-qubit gates U_j are gate independent. That is, every layer of one-qubit gate suffers the same noise.

The first assumption is made without loss of generality, since most quantum circuits can be recompiled in the required form. The second and third assumptions are standard in the literature and allow us to encompass a broad class of noise and error processes; notably, requiring that the noise is CPTP does not include errors such as leakage (see Sec. III F). Crucially, when combined together, these three assumptions enable us to use randomized compiling on the target circuit, that is, to transform arbitrary noise processes into Pauli noise. In the reminder of the subsection, we thus assume that every layer in $\mathcal{C}$ is subject to Pauli noise.

The trap circuits are generated by creating a copy of the target circuit, and by replacing every single-qubit gate in this copy with either I, H, or $S = \sqrt{Z}$. These extra gates are undone by compiling their inverses in the subsequent single-qubit gate layer. Note that by the third assumption, each trap generated in this way is affected by noise that is identical to that affecting the target circuit (i.e., gate-independent Pauli noise for single-qubit gates), as it is equal to the target circuit except for the individual single-qubit gate layers.

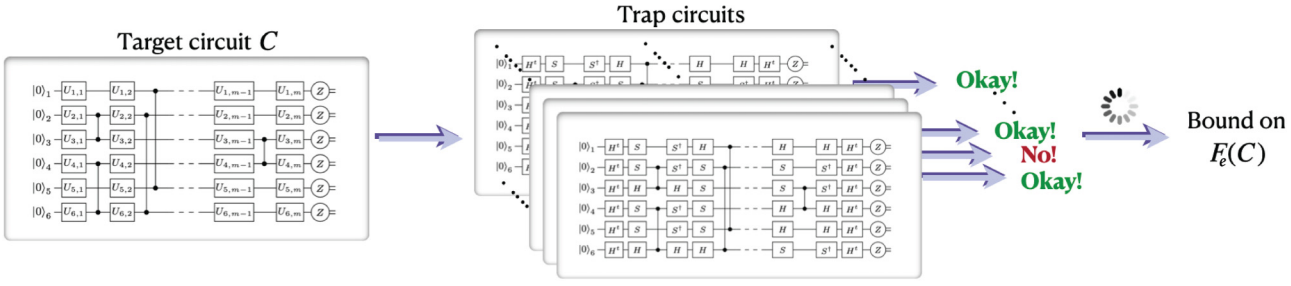


FIG. 42. Circuit output accreditation. Circuit output accreditation generates N_c circuits based on the target quantum circuit $\mathcal{C}$. These circuits (called *traps*) contain the same two-qubit gates as the target circuit, and after the neighboring layers of one-qubit gates are compiled into a single layer, every trap has the same width and depth as the target circuit. However, unlike the target circuit, in the absence of errors every trap returns a fixed, known output. By postprocessing the number of incorrect trap outputs, accreditation protocols return a lower bound on the process fidelity $F_e(\mathcal{C})$ of the target circuit.

In the absence of errors, the traps always return a fixed outcome $(0, 0, \ldots, 0)$. However, if an error occurs, the trap returns an incorrect output with probability larger than 50%. The proof of this statement (which is provided in detail in Ref. [409]) requires commuting errors all the way to the end of the circuit, and showing that due to the effect of the randomly chosen one-qubit gates, they have at least 50% probability of flipping one or more bits in the output string. Building on this property of the traps, the accreditation protocol takes the following steps:

(1) Generate and run a trap circuit. If the trap circuit returns the bit-string $(0, 0, \ldots, 0)$, mark the run as "successful." Otherwise, mark it as "unsuccessful."
(2) Repeat the step above a number $N_c = 2\ln(2/(1-\alpha))/\theta^2$ and calculate the total number $N_{\text{uns}} \in [0, N_c]$ of unsuccessful runs.

After all the traps have been run, the process fidelity $F_e(\mathcal{C})$ of the target quantum circuit $\mathcal{C}$ is bounded above and below by N_{uns} up to an error $\mathcal{O}(\theta)$:

$$1 - \frac{N_{\text{uns}}}{N_c} \geq F_e(\mathcal{C}) \geq 1 - 2\frac{N_{\text{uns}}}{N_c}. \tag{420}$$

Thus, circuit accreditation enables lower and upper bounding the circuit fidelity.

XI. HOLISTIC BENCHMARKS

Holistic benchmarks are methods for quantifying the overall performance of a quantum computer. These methods typically summarize important aspects of a quantum computer's performance in relatively few numbers or plots, such as the quantum volume [137] or capability regions [294]. Most holistic benchmarks quantify the impact of errors on overall performance, but they typically do not directly quantify gate (or layer) error rates, unlike RB protocols (Sec. VIII). Holistic benchmarks, therefore, complement and contrast with both detailed error characterization tools like tomography (Sec. VII) and RB protocols (Sec. VIII). In this section, we discuss some of the most widely used or important holistic benchmarking methods. We discuss the following areas within holistic benchmarking:

(a) *Volumetric benchmarks* (Sec. XI A). Volumetric benchmarking [410] is a framework that encompasses many different benchmarks. We discuss this framework, and two of its specific benchmarks or benchmark families: the quantum volume benchmark [137] and mirror circuit benchmarks [294].
(b) *Application benchmarks* (Sec. XI B). Holistic benchmarks based on applications or algorithms are now widely used to benchmark and compare quantum computers. We overview some of these methods, using examples from two algorithmic benchmarking suites.
(c) *Scalable holistic benchmarks* (Sec. XI C). Many existing holistic benchmarks are not scalable, but there are now techniques for creating scalable benchmarks from any set of circuits or algorithms. We briefly discuss these methods.

A. Volumetric benchmarks

Volumetric benchmarking [410] is a general methodology for benchmarking, rather than a specific benchmark. It generalizes ideas first introduced in the quantum volume benchmark (Sec. XI A 1). Volumetric benchmarks quantify a quantum computer's ability to run circuits with low error. In contrast to randomized benchmarks, volumetric benchmarks do not *directly* quantify the error rates of a quantum computer's qubits or gates. Instead, they quantify a quantum computer's rate of errors when running circuits of various shapes. A specific volumetric benchmark is defined by the following.

(1) A circuit family $\mathcal{C}_{W,D} = \{C\}$, that is indexed by circuit width (W, i.e., the number of qubits) and circuit depth (D). Note that "circuit depth" need not refer to the total number of layers of native gates in a low-level circuit; instead, for example, it could refer to the number of n-qubit Clifford gates in the circuit or the number of repetitions of an algorithmic subroutine (see discussion in Ref. [410]).
(2) A method for selecting circuits from $\mathcal{C}_{W,D}$ for each circuit shape (W, D) (e.g., a probability distribution over $\mathcal{C}_{W,D}$ for each W and D).
(3) An error metric or measure of success (e.g., total variation distance, classical fidelity, etc.; see Sec. IV) with which to compute how well any circuit in $\mathcal{C}_{W,D}$ was performed on a quantum computer.

Examples of circuit families for which a volumetric benchmark can be defined include the quantum volume circuits (see Fig. 45), randomized mirror circuits (see Sec. XI A 2), or the circuits from many algorithms.

Applying a volumetric benchmark to a quantum computer consists of the following.

(1) Picking a range of circuit shapes (W, D) at which to run circuits.
(2) At each chosen (W, D), selecting circuits from $\mathcal{C}_{W,D}$ and running them (the kinds of permissible compilation rules for each circuit depends on the benchmark).
(3) From each circuit's data, estimating how well the quantum computer ran that circuit using the benchmark's performance metric.

This procedure generates data consisting of the quantum computer's performance on the benchmark's circuit family as a function of both width and depth. That data is then typically displayed on the width × depth plane—sometimes called a "volumetric benchmarking plot"—as demonstrated in Fig. 43. Such a plot is a high-level overview of a quantum computer's performance on the circuits from that benchmark's circuit family. Volumetric plots can also be used to informally assess the kinds of errors occurring in the benchmarked system—e.g., crosstalk errors cause circuit error rates to increase faster with increasing circuit width than would be predicted by gate error rates measured using isolated one- and two-qubit RB [294].

The volumetric plot in Fig. 43 is a high-level performance summary, but it still contains a lot of detail. Therefore, it is sometimes useful to provide more concise and easily understood performance summaries. One way to do this is with "capability regions" [294]. Capability regions use volumetric benchmarking data to compute regions where a quantum computer can and cannot successfully run circuits, using some threshold for "success." A capability region constructed from the data of Fig. 43 is shown in Fig. 44.

1. Quantum volume

The *quantum volume (QV)* benchmark [137] is a holistic benchmark that inspired volumetric benchmarking. The QV benchmark runs randomly sampled "square" circuits with a particular structure, and computes a single number—the "quantum volume"—summarizing a system's performance on those circuits. The QV benchmark explicitly permits compilation of its circuits, so it jointly tests a quantum computing system's compilers and gates, i.e., it is a "full-stack" benchmark [137,317,411]. QV is currently one of the most widely used metrics for comparing integrated quantum computing systems, and summarizing the field's overall progress (e.g., see Ref. [412]). Note, however, that as of 2025 IBM (the developers of QV) are primarily using the error per layered gate (EPLG) (see Sec. VIII E) to quantify their systems' overall performance.

The QV benchmark is based on the circuits shown in Fig. 45, referred to as *randomized model circuits* or simply *quantum volume circuits*. The QV circuits are defined for any shape (W, D), but the QV benchmark uses only circuits of this type that are "square" (i.e., have equal width and depth). Each layer in an n-qubit QV circuit comprises $n/2$ disjoint two-qubit gates, between random pairs of qubits (where $n/2$ is rounded down if n is odd, and the single qubit that is not in any pair idles). Each gate is a uniformly random two-qubit unitary [i.e., it is drawn from the Haar measure on SU(4)]. The QV analysis (described below) uses the concept of the *heavy outputs* of a probability distribution (see Sec. IV A 4). The *heavy* outputs are the half of

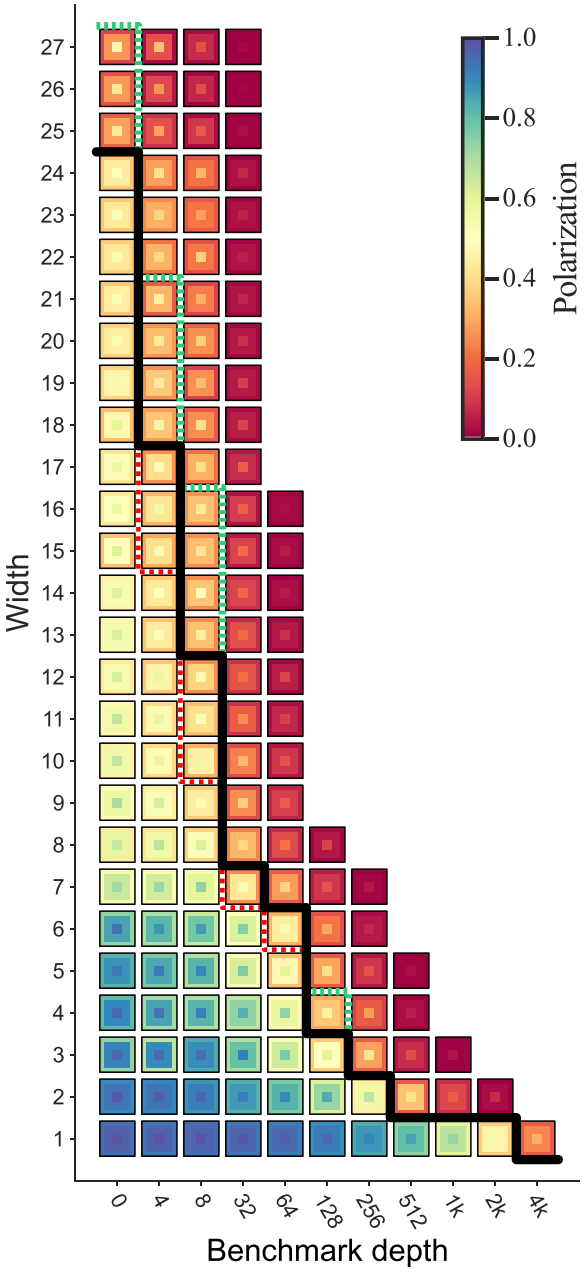


FIG. 43. Volumetric benchmarking. The results of a volumetric benchmark run on `ibmq_montreal`. This benchmark consists of running randomized mirror circuits (Sec. XI A 2) of various shapes. For each circuit width and benchmark depth (see main text), the concentric squares show the maximum (inner square), mean (middle square), and minimum (outer square) of the estimated polarizations S_{pol} [Eq. (350)] for all the circuits of that shape that were run. Frontiers (green, black, and red lines) show the circuit shapes at which these three statistics drop below the threshold value of $1/e$. (Figure reprinted with permission from [340].)

the outputs that are most likely to appear, i.e., those whose probability is above the median probability. For example, if the bit strings 00, 01, 10, and 11 have probabilities 0.1,

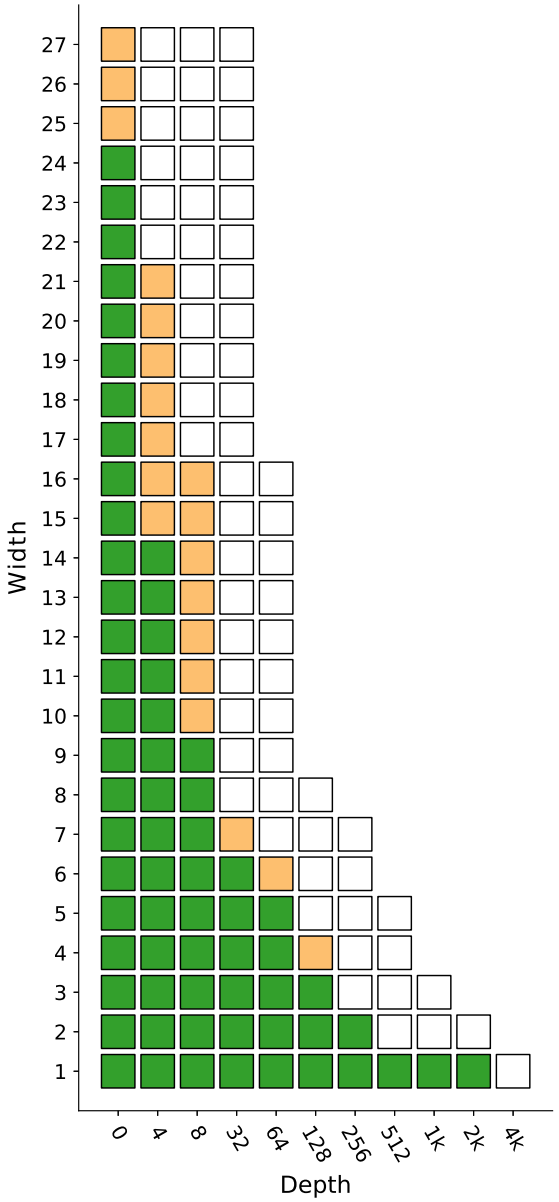


FIG. 44. Capability regions. Capability regions [294] are high-level summaries of a quantum computer's ability to execute circuits with low error. This plot is a capability region created from the volumetric benchmarking data of Fig. 43, and it summarizes the circuit shapes at which all randomized mirror circuits that were executed in this experiment ran successfully. Here, "success" means that a circuit's polarization [Eq. (350)] is above the threshold of $1/e$. A square is green if all circuits ran successfully, orange if some succeeded and some failed, and white if no circuits succeeded. Capability regions will typically depend on the circuit family used to construct them, e.g., a circuit family with a higher two-qubit gate density will typically result in a smaller green region.

0.2, 0.3, and 0.4, respectively, the heavy outputs are 10 and 11.

The QV benchmark consists of applying the following "quantum volume test" for increasingly large n:

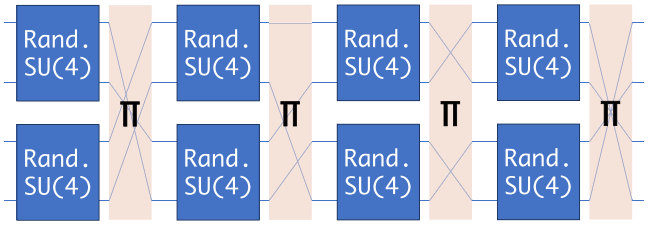


FIG. 45. Quantum volume circuits. A depth D (here $D = 4$) quantum volume circuit on W qubits (here $W = 4$). Each layer consists of Haar-random two-qubit unitaries on a random pairing of the W qubits (denoted here by random two-qubit gates between neighboring qubits and a random permutation of all the qubits). The quantum volume benchmark uses only square circuits, i.e., $D = W$. When W is odd, one randomly selected qubit idles during each layer.

(1) Sample shape (n, n) QV circuits.
(2) For each sampled circuit, compile it into a circuit that can be run on the specific system being tested. Approximate compilations are permissible (trading off fewer gates, and their associated errors, for intrinsic synthesis error in the circuit), but a faithful attempt to approximately implement each circuit's unitary is required.
(3) Run each compiled circuit many times, and for each circuit estimate the probability h of a heavy output (this is estimated by simply computing the observed frequency of heavy outputs).
(4) Assess whether $\bar{h} > 2/3$ with 95% confidence, where $\bar{h}$ is h averaged over all sampled circuits of shape (n, n). If $\bar{h} > 2/3$ with 95% confidence, then the system passes the n-qubit QV test, and otherwise it fails.

A system's QV is 2^n, where $n + 1$ is the smallest integer for which the system fails the QV test. For instance, if a six-qubit QV test achieves $\bar{h} > 2/3$, but a seven-qubit QV test achieves $\bar{h} < 2/3$, the measured QV is $2^6 = 64$.

The QV threshold of $\bar{h} > 2/3$ is somewhat arbitrary, but it can be motivated as follows. In the absence of errors, deep and wide QV circuits have $h \approx (1 + \ln 2)/2 \approx 0.85$. In the presence of errors that completely depolarize all of the qubits by the end of a QV circuit—i.e., the qubits are in the maximally mixed state by the end of the circuit—then $h = 0.5$. The threshold value of $2/3$ is approximately half way between these two regimes, which corresponds to a probability of an error in the compiled QV circuits of approximately 50%.

The QV benchmark favors quantum computers with high connectivity and gate set expressivity. For example, if one n-qubit system has linear connectivity and another has all-to-all connectivity, but they both have the same error rates on their one- and two-qubit gates, the all-to-all connectivity device will (typically) have a significantly

higher QV. A benchmark that favors higher connectivity is reasonable, but it is not a universally good choice. Higher connectivity is likely broadly useful for NISQ algorithms, but is not useful under all circumstances (e.g., it is not needed to run quantum error correction using surface codes). Alternative versions of the QV benchmark with lower connectivity in the random circuits are possible [411].

The QV benchmark is not scalable, because estimating h requires the simulation of the QV circuits to compute the heavy outputs, which is exponentially expensive. However, scalable adaptations of this benchmark have been proposed in Refs. [317,411].

2. Mirror circuit benchmarks

Mirror circuit benchmarks [294] are a family of volumetric benchmarks based on mirror circuits. Mirror circuits [see Fig. 46(a)] are a form of motion-reversal circuit that are constructed by (1) following a circuit ($\mathcal{C}$) with its layer-by-layer inverse ($\mathcal{C}^{-1}$), (2) adding in random Pauli gates between these two circuits, to prevent systematic error cancellation or addition between the two halves of the circuit, and (3) randomizing the state preparation and measurement basis of each qubit. Unlike the QV benchmark, mirror circuit benchmarks are *not* full-stack benchmarks (although mirror circuit benchmarks can be adapted to full-stack benchmarking [411]). This is because, like RB circuits, mirror circuits must not be arbitrarily compiled, as each mirror circuit's overall operation is simply bit flips on some of the qubits. Instead, mirror circuit benchmarks are designed to measure a system's ability to implement low-level circuits, complementing metrics that also incorporate compiler performance, like the QV.

Mirror circuits can be used to construct scalable benchmarks from any set of circuits (see Sec. XI C). Here, we discuss two specific mirror circuit benchmarks introduced in Ref. [294]: *randomized mirror circuit* and *periodic mirror circuit* benchmarks, shown in Figs. 46(b) and 46(c), respectively. Randomized mirror circuits are the same circuits that are used in mirror randomized benchmarking (MRB) (see Sec. VIII C 3) and averaged circuit eigenvalue sampling (ACES) (see Sec. IX C 2). Volumetric benchmarking with randomized mirror circuits is a scalable way to assess performance of a quantum computer on random, unstructured circuits. Like other random circuits [e.g., RB or cross-entropy benchmarking (XEB) circuits], these circuits scramble errors. Therefore, two different but equal-shape randomized mirror circuits typically have fairly similar performance, particularly as both circuit width and depth increases. In contrast, periodic mirror circuits are extremely ordered: they consist of repeating a short n-qubit "germ" circuit (that is randomly sampled from a distribution over possible short germ circuits). These circuits are not scrambling, but instead amplify

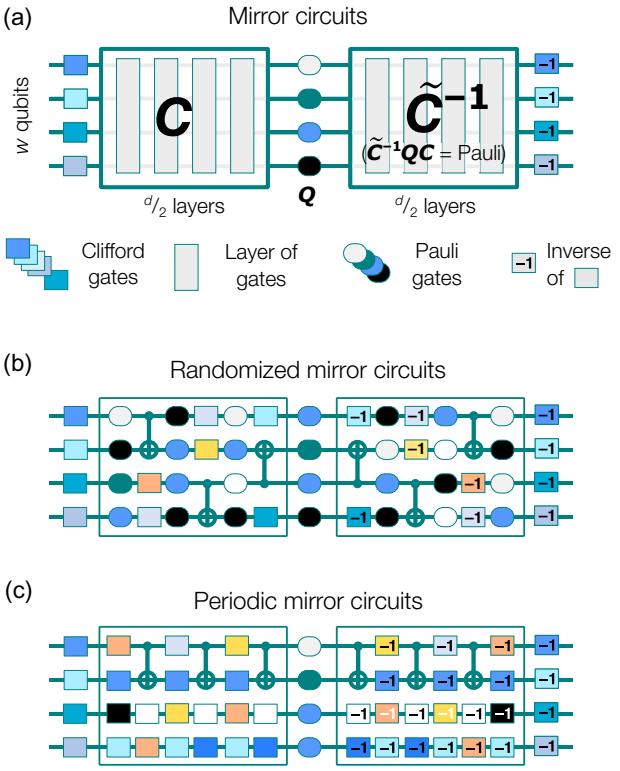


FIG. 46. Mirror circuit benchmarks. (a) Mirror circuits are a family of circuits with a motion reversal structure that are designed to enable scalable benchmarking. Two kinds of mirror circuit are (b) randomized mirror circuits and (c) periodic mirror circuits [294].

particular errors—with the particular errors that are amplified depending on the germ, as in long-sequence gate set tomography (see Sec. VII D 2). Therefore, in the presence of structured errors, such as coherent errors or biased stochastic Pauli errors, the variance in performance on periodic circuits will typically be much higher than with random circuits. Figure 47 shows how the performance of one system (`ibmq_london`) differs on random and periodic mirror circuits, illustrating how these benchmarks can be used to reveal structured errors.

B. Application benchmarks

Benchmarks based on algorithms and applications can be used to quantify the performance of quantum computing systems. Many different benchmarks fall under the category of "application benchmarks," which encapsulates both high-level applications, such as solving a MaxCut problem or finding a ground state, variational algorithms such as the quantum approximate optimization algorithm (QAOA), as well as key subroutines, such as the quantum Fourier transform (QFT) or quantum error correction. The primary purpose of application benchmarks is to measure the performance of a full-stack quantum computer for a specific use case or algorithm. This contrasts with most

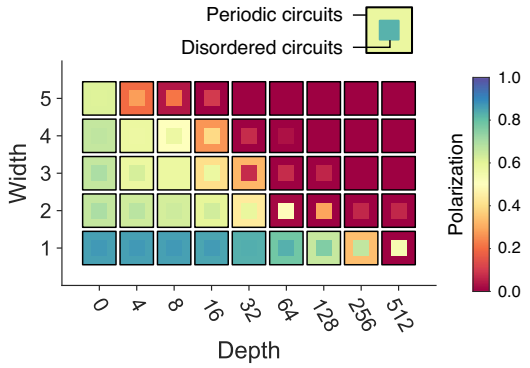


FIG. 47. Periodic and random circuit benchmarks. A volumetric plot comparing a quantum computer's performance on disordered and highly structured circuits. This plot shows the polarization of the worst-performing circuit over 40 disordered and periodic circuits (randomized mirror circuits and periodic mirror circuits, respectively). Worse performance on periodic circuits (as seen here) is a signature of structured errors, such as coherent errors. Plot created using `ibmq_london` data from Ref. [294].

characterization and other benchmarking protocols, which aim to measure specific properties (e.g., qubit coherence or gate fidelities) of low-level components.

Numerous application-oriented quantum benchmarking suites have been recently developed, including `SupermarQ` [413], `QASMBench` [414], and those developed by the `QED-C` [415,416]. These various libraries are based on similar ideas, cover a range of application domains, and can be implemented on various quantum computing architectures (e.g., gate-based quantum computers, quantum annealers, etc.). In this subsection, we review several examples of application benchmarks from the `SupermarQ` and `QED-C` suites. However, because application benchmarking encompasses many diverse methodologies, the reader is encouraged to review other works for a broader perspective of the entire field of application benchmarks [413–418].

In Fig. 48, we show example circuits for three different benchmarks: `SupermarQ`'s `Vanilla QAOA`, `SupermarQ`'s `Phase Code`, and the `QED-C` suite's `QFT(1)`. Both the `Vanilla QAOA` and `Phase Code` benchmarks are examples of *proxy* applications, which focus on a specific aspect of a larger, end-to-end application. For example, the `Vanilla QAOA` benchmark measures how well a quantum computer is able to execute a single instance of a variational circuit [419], whereas full execution of the standard QAOA algorithm would involve iterating over a large number of circuit instances (and classical optimization). Similarly, the `Phase Code` benchmark tests a quantum computer's ability to execute circuits containing mid-circuit measurements—which tests a component used in, e.g., syndrome extraction—but does not use the results of these measurements to correct errors in the circuit. The `QFT(1)` benchmark, which contains both the QFT and its inverse, is an example of a key subroutine that appears in many quantum algorithms including Shor's algorithm [21] and the HHL algorithm [23]. Application benchmarks should specify the compiler optimizations that may be applied to the circuits prior to their execution.

Figure 49 shows the results of running application benchmarks from the `SupermarQ` and `QED-C` suites on `ibmq_guadalupe`, represented as a volumetric plot. Each colored rectangle corresponds to

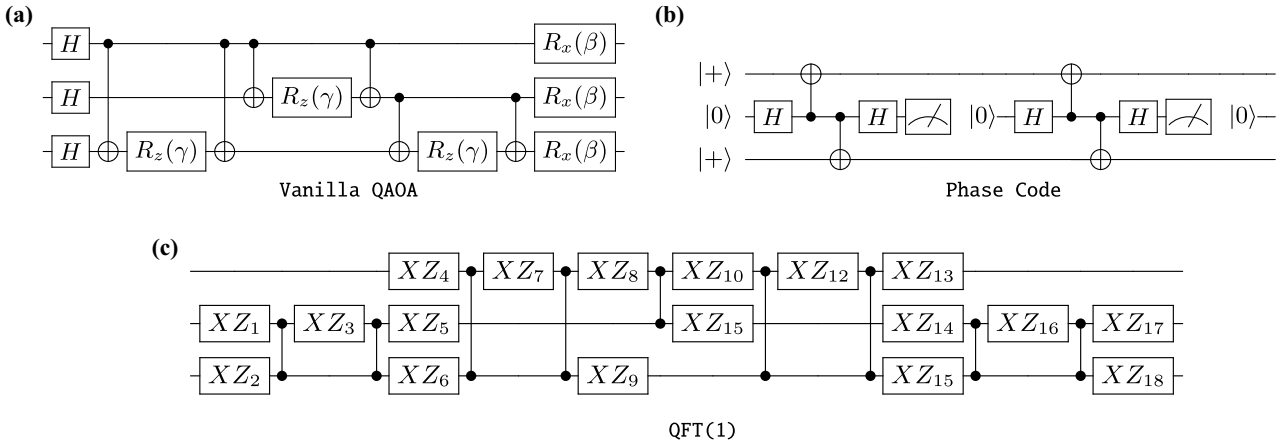


FIG. 48. Example application benchmarks. (a) Three-qubit `Vanilla QAOA` benchmark from `SupermarQ`. In general, the γ and β parameters need to be optimized with respect to a specific objective function (e.g., a Hamiltonian ground state) in a variational quantum-classical loop. However, this benchmark can be designed to be measured for a single instance of γ and β to minimize errors due to drift (e.g., in cloud-based systems). (b) Two-qubit `Phase Code` benchmark from `SupermarQ`. The phase code utilizes an ancilla qubit [middle] to detect phase flips on the data qubits [outer] using mid-circuit measurements. The ancilla qubit must be reset for each cycle of error detection. (c) Three-qubit `QFT(1)` benchmark from `QED-C` suite. The secret integer for this benchmark is $x = 2$, and the XZ_i gates are `PhasedXZ` gates where $XZ_i(a,x,z) = Z^z Z^a X^x Z^{-a}$.

an individual benchmark with the color indicating the score achieved. The benchmark score is a value ranging from 0 (poor performance) to 1 (best performance). The specific definition of the score function depends on the benchmark (see, e.g., the benchmark definitions given in [413] and [415]), and typical examples include evaluating expectation values or computing the classical (Hellinger) fidelity [Eq. (178)]. Application benchmarks are now sometimes used to make high-level comparisons between different quantum computing systems [413–416,420].

One useful aspect of application benchmarks is that their circuits typically have diverse properties, and so they may stress the quantum computer in diverse ways. This contrasts with benchmarks based on random circuits (e.g., RB protocols and the QV), which contain similar circuit structures. Initial efforts to profile quantum programs have underscored the distinct differences between applications originating from domains such as quantum chemistry and combinatorial optimization [413,414]. Figure 50 shows how the circuits of four different application benchmarks have significantly different properties [413]. It does so by plotting the values for six different features for each circuit: *program connectivity* (PC), *parallelism* (Par), *measurement* (Mea), *liveness* (Liv), *entanglement-ratio* (Ent), and *critical depth* (CD). The program connectivity of an n-qubit circuit is computed as $\sum_i^n \mathcal{D}(q_i)/(n^2-n)$, where $\mathcal{D}(q_i)$ is the degree of qubit q_i in the program's connectivity graph. This gives the program connectivity a range between zero, for programs without entangling gates, to one for a program with a complete connectivity graph. The parallelism feature relates the total number of gates (N_G) and circuit depth (D) within the expression $(N_G - D)/(nD - D)$ to capture the amount of parallelism available within a quantum program. The parallel execution of gates often exhibit correlated crosstalk which degrades circuit performance. The measurement feature is given by L_{mcm}/D for a circuit with L_{mcm} layers containing at least one mid-circuit measurement operation, during which the idling spectator qubits can dephase. The liveness feature also considers the idling time of qubits. It is computed as $(\sum_{ij} A_{ij})/(nD)$, where A is a binary $(n \times D)$ matrix with entry $A_{ij} = 1$ if qubit i is acted on by a gate at time step j, otherwise the entry is zero. Entanglement-ratio is given by the fraction of entangling gates divided by the total gate count, and it can provide insights into program behavior if the specific hardware running the benchmark has large differences in one- and two-qubit gate error rates. Finally, the critical-depth feature is computed by counting the number of entangling gates along the program's critical path and dividing by the total number of entangling gates in the circuit.

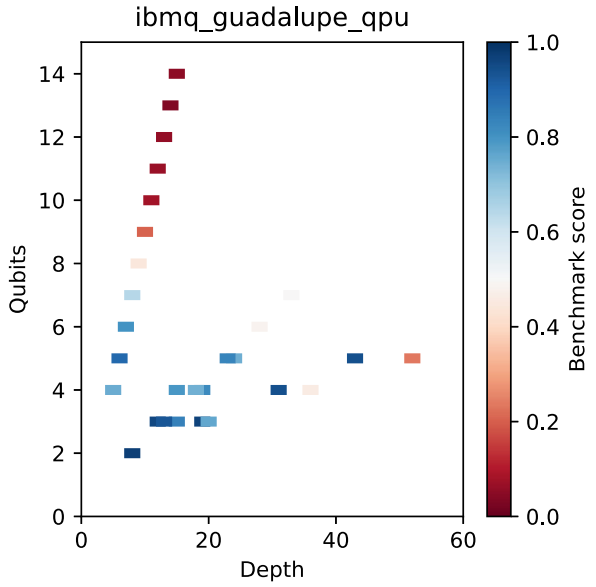


FIG. 49. Volumetric benchmarking with application circuits. Evaluation of `SupermarQ` and `QED-C` benchmarks on a superconducting quantum computer (`ibmq_quadalupe`) for a variety of different circuit widths and depths. Each point represents the execution of a single application benchmark, where the color corresponds to the "score" achieved for that benchmark. Note that these benchmarks include the circuit compilation process in their evaluation (i.e., "depth" refers to the circuit depth of the uncompiled circuit). The exact expression used to obtain the score varies from benchmark to benchmark and can be found in the definitions of the benchmark suites [413,415]. Examples of a benchmark score include the measurement of an expectation value or classical fidelity. We observe that the performance falls off with circuit width; this could be because this system has limited connectivity, necessitating SWAP gates to implement nonlocal two-qubit gates.

Each feature is meant to capture some salient aspect of a quantum program. These features, combined with the benchmark results in Fig. 49, can be used to correlate system performance with program profiles (see Fig. 50). Each square in Fig. 50 corresponds to the coefficient of determination (R^2) for a particular pair of device and program feature. In other words, it shows the proportion of the variation in that device's performance, across all of the evaluated application benchmarks, which can be explained by that particular program feature. Each R^2 value is obtained by performing a linear regression over all of the device's scores on the benchmarks (dependent variable) and the specific values of the particular program feature for each benchmark (independent variable).

C. Scalable holistic benchmarks

Benchmarking a quantum computer's performance on large circuits or applications is an inherently difficult task because it is not typically feasible to compute what the correct outcome should be using simulations on a classical computer. Therefore, many existing holistic benchmarks

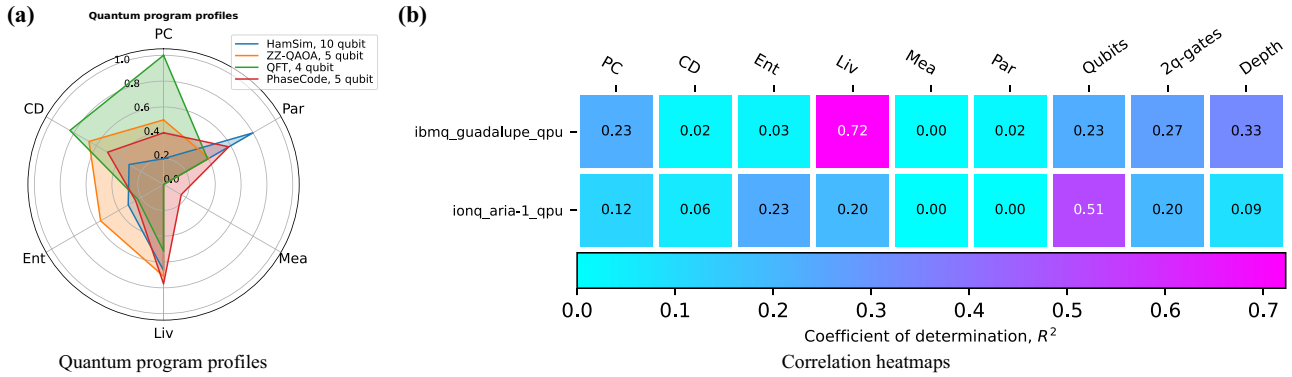


FIG. 50. Profiling application benchmarks. (a) Quantum program profiles. Application benchmarks have many diverse properties that allow one to probe different properties of a quantum processor. We quantify the program connectivity (PC), parallelism (Par), measurement (Mea), liveness (Liv), entanglement-ratio (Ent), and critical depth (CD) for the 10-qubit Hamiltonian simulation (HamSim) benchmark, the 5-qubit ZZ-QAOA benchmark, the 4-qubit QFT circuit, and the 5-qubit phase code [413]. (b) Correlation heatmaps. After running the application benchmarks on a quantum computer, the observed performance can be correlated with the program features to produce the above heatmap. For example, the performance of the superconducting device, with relatively short coherence times, shows a high correlation (0.72) with the liveness feature of the benchmark circuits.

either do not scale beyond around 50 qubits (examples include the QV benchmark and XEB used for demonstrations of "quantum supremacy") or they only use circuits from some restricted circuit class that can be efficiently simulated classically (e.g., Clifford circuits, as in many RB methods). In the context of application benchmarks, the applications or circuits used are often designed to circumvent this "verification" problem, by ensuring that a quantum computer's performance on the benchmark can be quantified without inefficient circuit simulations [413]. Often this is achieved with algorithm-specific methods, e.g., by creating a benchmark that tests a quantum computer's ability to solve the one-dimensional transverse field Ising model [421], or to prepare easy-to-verify states such as a GHZ state [422], or to measure operators with a known upper bound (as is done in the Mermin-Bell benchmark [413]).

A complementary approach to creating scalable holistic benchmarks is to (1) choose a benchmark's circuits without addressing the efficiency problem, and then (2) measuring a quantum computer's performance on those circuits *indirectly*. This can be achieved with any technique that can efficiently estimate a quantum computer's circuit execution fidelity (or some other interesting metric of circuit performance) for an arbitrary circuit. We discussed two such techniques—mirror circuit fidelity estimation and circuit output accreditation—in Sec. X. This can even enable scalable full-stack benchmarks, as discussed in Ref. [411].

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research Quantum Testbed Program under Contract No. DE-AC02-05CH11231 and DE-SC0021526, as well as the Quantum Testbed Pathfinder Program. A.H. acknowledges financial support from the Berkeley Initiative for Computational Transformation Fellows Program. T.P. acknowledges support from an Office of Advanced Scientific Computing Research Early Career Award. L.J. and S.C. acknowledge support from the ARO (W911NF-23-1-0077), ARO MURI (W911NF-21-1-0325), AFOSR MURI (FA9550-19-1-0399, FA9550-21-1-0209, FA9550-23-1-0338), DARPA (HR0011-24-9-0359, HR0011-24-9-0361), NSF (OMA-1936118, ERC-1941583, OMA-2137642, OSI-2326767, CCF-2312755), NTT Research, Packard Foundation (2020-71479). A.H. acknowledges fruitful discussions with Joel J. Wallman, Joseph Emerson, Ian Hincks, and Arnaud Carignan-Dugas. We acknowledge useful feedback from Juan Jesus Gonzalez De Mendoza Prada and Robin Harper. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide

public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

All authors contributed to the writing of the manuscript.

All authors declare no competing interests.

DATA AVAILABILITY

All data are available from the corresponding author upon reasonable request.

APPENDIX A: ERROR GENERATORS

In Sec. II C, we introduced various representations for modeling errors acting on quantum processes. These representations, including Kraus operators, transfer matrices, process matrices, and Choi matrices, are able to capture arbitrary CPTP gate errors. Unfortunately, it can be difficult to tease apart an arbitrary CPTP map and relate components of observed error matrices to known error sources, such as qubit fluctuations or systematic calibration errors in gates. A more immediate connection can be made using *error generators*. Error generators are designed to capture and categorize those *small* Markovian errors in quantum gates that appear in reasonably well-behaved quantum computers. In what follows, we denote the ideal transfer matrix of a gate G to be Λ_G. Now, using the composition property of transfer matrices (see Sec. II C 2), we may write the transfer matrix of the noisy quantum gate $\tilde{\Lambda}_G$ as

$$\tilde{\Lambda}_G = \Lambda_{\mathcal{E}}\Lambda_G\,, \tag{A1}$$

where $\Lambda_{\mathcal{E}}$ is the transfer matrix (e.g., PTM) that captures the noise and errors impacting Λ_G [423]. If the error is small, then the noisy gate is close to the target unitary (i.e., $\tilde{\Lambda}_G \approx \Lambda_G$) and $||\Lambda_{\mathcal{E}} - \mathbb{I}|| \ll 1$. By taking the $\log(\Lambda_{\mathcal{E}})$, we can learn how much $\Lambda_{\mathcal{E}}$ deviates from $\mathbb{I}$, or rather how much $\tilde{\Lambda}_G$ deviates from Λ_G. Thus, we define the error generator [45] of $\Lambda_{\mathcal{E}}$ to be

$$\mathcal{L} = \log(\Lambda_{\mathcal{E}}) \simeq \Lambda_{\mathcal{E}} - \mathbb{I}\,, \tag{A2}$$

such that we may write Eq. (A1) as

$$\tilde{\Lambda}_G = e^{\mathcal{L}}\Lambda_G\,. \tag{A3}$$

Here, $\mathcal{L}$ is the generator of $\Lambda_{\mathcal{E}}$ in analogy with how Hamiltonians are the generators of unitary transformations.

In order to ensure that $\mathcal{L}$ generates a CPTP map, it must be expressible as a Lindblad superoperator [424]. Expanding the Lindblad equation in a basis of Pauli operators $\{P_j\}$ we have

$$\mathcal{L}(\rho) = \sum_j \epsilon_j [P_j, \rho] + \sum_{j,k} h_{j,k}\left(P_j\rho P_k\rho P_j - \frac{1}{2}\left\{P_j^\dagger P_k, \rho\right\}\right). \tag{A4}$$

Here, $\epsilon_j \in \mathbb{R}$ characterizes the size of unitary errors, and $h_{j,k}$ is positive semidefinite and quantifies the type and rate of the dissipative dynamics.

While the (H)amiltonian error rates (ϵ_j) that describe the unitary dynamics in Eq. (A4) are often relatively easy to understand (e.g., a Pauli-X error on an X gate corresponds to an over or under rotation), the dissipative part can be more challenging. So, the error generator framework splits those errors into symmetric (or *stochastic*) and antisymmetric (or *active*) components. The (A)ctive components can be derived from couplings to quantum degrees of freedom and are responsible for, e.g., nonunital effects such as amplitude damping. The stochastic components are precisely those that might arise from fluctuating Hamiltonian terms. Because stochastic *Pauli* errors appear so frequently (e.g., in Pauli frame randomization, randomized compiling, and models for quantum error correction), the symmetric sector is further decomposed into a Pauli (S)tochastic and stochastic (C)orrelation sectors corresponding to the diagonal and off-diagonal symmetric terms. Thus, we define the following elementary generators $\{H, S, C, A\}$:

$$H_P(\rho) = -i[P, \rho]\,, \tag{A5}$$

$$S_P(\rho) = P\rho P - \rho\,, \tag{A6}$$

$$C_{P,Q}(\rho) = P\rho Q + Q\rho P - \frac{1}{2}\{\{P, Q\}, \rho\}\,, \tag{A7}$$

$$A_{P,Q}(\rho) = i\left(P\rho Q - Q\rho P + \frac{1}{2}\{[P, Q], \rho\}\right). \tag{A8}$$

They form a complete basis for superoperators. An error generator $\mathcal{L}$ is a Lindbladian superoperator that generates coherent, stochastic, and/or nonunital gate errors. Therefore, we may write $\mathcal{L}$ as a linear combination of elementary error generators,

$$\mathcal{L} = \mathcal{L}_{\mathbb{H}} + \mathcal{L}_{\mathbb{S}} + \mathcal{L}_{\mathbb{C}} + \mathcal{L}_{\mathbb{A}}\,, \tag{A9}$$

$$= \sum_P h_P H_P + \sum_P s_P S_P\,, \tag{A10}$$

$$+ \sum_{P,Q>P} c_{P,Q}C_{P,Q} + \sum_{P,Q>P} a_{P,Q}A_{P,Q}\,, \tag{A11}$$

where the coefficients $\{h, s, c, a\}$ denote the *error rates* of each error process, and where all errors and their corresponding rates are indexed by one (or two) distinct Pauli

operators P (and Q). Thus, any arbitrary error generator $\mathcal{L}$ can be written as a linear combination of all of the elementary error generators. This framework enables one to quantify the rates of different errors afflicting quantum gates [46].

APPENDIX B: GROUPS AND GATE SETS

1. Groups

A *group* $\mathbb{G}$ is a mathematical set of operational elements $\{G_i\}$ that satisfy the following basic properties:

(1) Closure: $\forall\, G_i, G_j \in \mathbb{G},\ G_i \cdot G_j = G_k \in \mathbb{G}$.
(2) Associativity: $(G_i \cdot G_j) \cdot G_k = G_i \cdot (G_j \cdot G_k)$.
(3) Identity element: $\exists\ \mathbb{I} \in \mathbb{G}$ s.t. $\mathbb{I} \cdot G_i = G_i \cdot \mathbb{I} = G_i$, $\forall\, G_i \in \mathbb{G}$.
(4) Inverse element: $\forall\ G_i \in \mathbb{G},\ \exists\ G_i^{-1} \in \mathbb{G}$ s.t. $G_i \cdot G_i^{-1} = G_i^{-1} \cdot G_i = \mathbb{I}$.

Below, we introduce some important groups and gate sets in quantum computing, and highlight the properties that distinguish each from the others.

2. The Pauli group

The n-qubit *Pauli group*, denoted $\mathbb{P}_n$, is the set of Pauli operators formed by the tensor product of all combinations of single-qubit Paulis for n qubits, multiplied by factors of ± 1 or $\pm i$:

$$\mathbb{P}_n = \{\pm 1, \pm i\} \times \{I, X, Y, Z\}^{\otimes n}, \tag{B1}$$

where

$$I = \sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{B2}$$

$$X = \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{B3}$$

$$Y = \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \tag{B4}$$

$$Z = \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{B5}$$

are the single-qubit Pauli operators. The n-qubit Pauli operators have a number of helpful properties, namely,

(1) they form a projective group under matrix multiplication,
(2) they are unitary and Hermitian,
(3) they are a trace-orthogonal basis for the space of operators,
(4) they correspond to natural Hamiltonians, and
(5) they form a unitary 1-design in $d = 2^n$-dimensional Hilbert space (see Sec. C 2).

TABLE V. $\mathbb{P}_1$ under H conjugation.

P	$HPH^\dagger$
I	I
X	Z
Y	$-Y$
Z	X

3. The Clifford group

The n-qubit *Clifford group* [425], denoted $\mathbb{C}_n$, is the set of operations that *normalize* the n-qubit Pauli group. This means that any element from the Clifford group $C \in \mathbb{C}_n$ maps each Pauli operator to another Pauli operator under conjugation:

$$\forall\, C \in \mathbb{C}_n : CPC^\dagger \mapsto P' \in \mathbb{P}_n, \quad \forall\, P \in \mathbb{P}_n\,. \tag{B6}$$

Typical examples of Clifford gates which are not in the Pauli group are the Hadamard H, $S = \sqrt{Z}$, CNOT, SWAP, and iSWAP gates. In fact, the subgroup of Clifford gates $\{H, S, \text{CNOT}\}$ is sufficient to generate the full Clifford group between any pair of qubits. In Tables V and VI, we show the action of all single-qubit Pauli operators under conjugation by the Hadamard H and S gates, respectively. In Tables VII and VIII, we show the action of all two-qubit Pauli operators under conjugation by the CNOT and iSWAP gates, respectively. In all cases, we find that the resulting gate is a Pauli, which belongs to $\mathbb{P}_n$.

The single-qubit Clifford group $\mathbb{C}_1$ contains 24 single-qubit gates; these include any integer number of $\pi/2$ rotations about any of the six cardinal axes of the Bloch sphere ($\pm\hat{x}$, $\pm\hat{y}$, and $\pm\hat{z}$), which includes all single-qubit Pauli gates ($\mathbb{P}_n \subset \mathbb{C}_n$). The size of the n-qubit Clifford group is given by [426]

$$|\mathbb{C}_n| = 2^{n^2+2n} \prod_{j=1,n} 4^j - 1\,. \tag{B7}$$

For example, the two-qubit Clifford group contains 11 520 elements, the three-qubit Clifford group contains 92 897 280 elements, the four-qubit Clifford group contains 12 128 668 876 800 elements, etc.

The Clifford group holds a special place in quantum computing. According to the *Gottesman-Knill theorem* [286], quantum circuits containing only Clifford gates and Pauli basis measurements can be efficiently simulated in

TABLE VI. $\mathbb{P}_1$ under S conjugation.

P	$SPS^\dagger$
I	I
X	Y
Y	$-X$
Z	Z

TABLE VII. $\mathbb{P}_2$ under CNOT conjugation.

P	CNOT(P)CNOT†
$I \otimes I$	$I \otimes I$
$I \otimes X$	$I \otimes X$
$I \otimes Y$	$Z \otimes Y$
$I \otimes Z$	$Z \otimes Z$
$X \otimes I$	$X \otimes X$
$X \otimes X$	$X \otimes I$
$X \otimes Y$	$Y \otimes Z$
$X \otimes Z$	$-Y \otimes Y$
$Y \otimes I$	$Y \otimes X$
$Y \otimes X$	$Y \otimes I$
$Y \otimes Y$	$-X \otimes Z$
$Y \otimes Z$	$X \otimes Y$
$Z \otimes I$	$Z \otimes I$
$Z \otimes X$	$Z \otimes X$
$Z \otimes Y$	$I \otimes Y$
$Z \otimes Z$	$I \otimes Z$

polynomial time on a classical computer. Therefore, Clifford circuits are insufficient to realize the full potential of quantum computers over classical computers. In fact, in order to perform universal quantum computation, one requires a gate set which also contains a non-Clifford gate, such as the $T = \sqrt{S}$ gate (sometimes called the "$\pi/8$" gate for historical reasons). Nonetheless, Clifford gates are ubiquitous in quantum computations and are essential to a number of important applications. For example, stabilizer codes in quantum error correction use Clifford gates for encoding and decoding. Additionally, benchmarking procedures for measuring average error rates of quantum gate sets, such as randomized benchmarking (Sec. VIII B), are constructed entirely of Clifford gates. Importantly, Clifford gates are used in these protocols because they form a unitary 2-design (and sometimes a unitary 3-design; see Sec. C 2).

TABLE VIII. $\mathbb{P}_2$ under iSWAP conjugation.

P	iSWAP$(P)i$SWAP†
$I \otimes I$	$I \otimes I$
$I \otimes X$	$Y \otimes Z$
$I \otimes Y$	$-X \otimes Z$
$I \otimes Z$	$Z \otimes I$
$X \otimes I$	$Z \otimes Y$
$X \otimes X$	$X \otimes X$
$X \otimes Y$	$Y \otimes X$
$X \otimes Z$	$I \otimes Y$
$Y \otimes I$	$-Z \otimes X$
$Y \otimes X$	$X \otimes Y$
$Y \otimes Y$	$Y \otimes Y$
$Y \otimes Z$	$-I \otimes X$
$Z \otimes I$	$I \otimes Z$
$Z \otimes X$	$Y \otimes I$
$Z \otimes Y$	$-X \otimes I$
$Z \otimes Z$	$Z \otimes Z$

4. Qudit groups and bases

The qubit Pauli group serves as a natural basis for analyzing qubit systems. However, no set of operators with the same properties exists for higher dimensional systems. Nevertheless, we can define two natural generalizations of the Pauli operators to higher dimensions, the *Weyl* and *Gell-Mann* operators that, taken together, satisfy all of the properties of the single-qubit Pauli group. In general, the Weyl operators allow one to more naturally generalize the machinery of qubit-based benchmarking routines. In contrast, the Gell-Mann matrices correspond more immediately to the underlying physical operations performed on a qudit based quantum processor, such as Rabi oscillations and Z gates in a two-level subspace of the qudit. In what follows, we refer to D as the dimension of the qudit, and $d = D^n$ as the dimension of the Hilbert space for n qudits.

a. The Gell-Mann basis

To construct the Gell-Mann operators, we can begin by embedding the single-qubit Pauli operators into two-dimensional subspaces of the higher-dimensional qudit space. Specifically, for a D-dimensional qudit, we can define these operators as

$$X^{jk} = |j\rangle\langle k| + |k\rangle\langle j|\,, \tag{B8}$$

$$Y^{jk} = i\,|j\rangle\langle k| - i\,|k\rangle\langle j|\,, \tag{B9}$$

$$Z^{jk} = |j\rangle\langle j| - |k\rangle\langle k|\,, \tag{B10}$$

where $0 \le j < k \le D$. We note that while the Z^{jk} are Hermitian operators, they are not linearly independent and thus cannot serve as a sufficient basis for qudit tomography. We can therefore extend the set $\{X^{jk}, Y^{jk} : 0 \le j < k < D\}$ to a trace-orthogonal basis with the inclusion of additional diagonal operators:

$$W^j = -j\,|j\rangle\langle j| + \sum_{0\le k<j} |k\rangle\langle k| \tag{B11}$$

for $1 \le j < D$. The combined set $\mathbb{G}_D = \{I, X^{jk}, Y^{jk} : 0 \le j < k < D\} \cup \{W^j : 1 \le j < D\}$ forms the *Gell-Mann basis* for qudit dimension D. Like the qubit Pauli matrices, every element of the Gell-Mann group is both traceless and Hermitian, and thus serves as a suitable choice for constructing qudit transfer matrices.

b. The Weyl group

Having constructed the qudit Gell-Mann basis, we now turn our attention to generalizing the qubit Pauli operators over the entire qudit space rather than embedding them in two-level subspaces of the qudit. This is known as the *Weyl group*, from which we can generalize qubit-based quantum algorithms and codes to qudits, as well as generalize the

Clifford group to higher dimensions. We begin by defining $\mathbb{Z}_D = \{0, 1, \ldots, D-1\}$, the additive group of the integers modulo D. Using $\mathbb{Z}_D$, we can generalize the qubit X and Z operators to the qudit space as

$$X = \sum_{j \in \mathbb{Z}_D} |j \oplus_D 1\rangle\langle j| \,, \tag{B12}$$

$$Z = \sum_{j \in \mathbb{Z}_D} \exp\left(\frac{2\pi i}{D} j\right) |j\rangle\langle j| \,, \tag{B13}$$

where $\oplus_D$ denotes addition modulo D. We note here that both X and Z compose to the identity under D rounds of self-multiplication, which is the natural generalization of the qubit X and Z operators squaring to the identity. The Weyl basis follows as a trace-orthogonal basis over $\mathbb{C}^{D\times D}$ defined as $\mathbb{W}_D = \{W_{xz} = X^x Z^z : x, z \in \mathbb{Z}_D\}$. Often in the literature, the qudit X operator is referred to as the "shift" operator as it increments the qudit state modulo D, and the qudit Z operator as the "clock" operator as it applies phases corresponding to multiples of the Dth root of unity. Finally, we can define the n-qudit Weyl group by taking the n-fold tensor product of all single-qudit Weyl matrices: $\mathbb{W}_{D,n} = \mathbb{W}_D^{\otimes n}$. It is worth noting that one can define multiqudit gates directly from the definitions of the Weyl matrices, similar to how the Hamiltonians for two-qubit gates are often defined in terms of n-qubit Paulis.

c. The n-qudit Clifford group

One important caveat about the Weyl basis is that, owing to the fact that it is not closed under multiplication, it is not a proper group. However, every element of the closure of the Weyl basis is related to an element of the Weyl basis up to an overall phase. We can therefore get rid of this overall phase by considering solely the adjoint action of the Weyl operators, and therefore defining a proper group we refer to as the "extended Weyl group," $\mathbb{EW}_D = \mathsf{U}(1)\mathbb{W}_D$ and the "extended n-qudit Weyl group" as $\mathbb{EW}_{D,n} = \mathsf{U}(1)\mathbb{W}_{D,n}$, where $\mathsf{U}(1)$ is the 1-dimension unitary group, allowing for arbitrary phases. Finally, with all this formalism defined, we can naturally define the n-qudit Clifford group to be the set of operators that normalize the extended Weyl group, or stated explicitly, the set $\mathbb{C}_{D,n}$ where

$$\mathbb{C}_{D,n} = \{U \in \mathsf{U}(D^n) : U\mathbb{EW}_{D,n}U^\dagger = \mathbb{EW}_{D,n}\} \,. \tag{B14}$$

APPENDIX C: RANDOMIZATION AND TWIRLING

The notion of *twirling* a quantum channel is a central component of many benchmarking methods based on randomized gate sampling. The basic concept of twirling is to average a quantum channel $\mathcal{E}$ over some unitary group. This allows one to measure the average performance of a quantum operation (e.g., a gate) for different combinations of input and output states, while reducing the resources needed for measuring the process fidelity of a gate compared to full quantum process tomography. As we will see in this section, twirling maps a dense CPTP matrix (modeled as a $d^2 \otimes d^2$ superoperator, where $d = 2^n$ for n qubits) into a block diagonal matrix, effectively condensing information about the physical process into the eigenvalues of the matrix. Below, we give a precise definition of twirling, and discuss the difference between twirling over the Pauli and Clifford groups.

1. The Haar measure

When twirling a quantum channel $\mathcal{E}$ over a unitary group in d dimensions, $\mathsf{U}(d)$, it is necessary to uniformly sample at random unitaries from $\mathsf{U}(d)$. The *uniform Haar measure*, denoted $\mu(U)$, is mathematical measure that is unique to each locally compact topological group which assigns equal weights to all elements of the group. $\mu(U)$ defines an integral over $\mathsf{U}(d)$ that is invariant under group transformations, and is normalized such that the total measure of the group is $\int d\mu = 1$. The uniform Haar measure defines how different elements of $\mathsf{U}(d)$ are weighted over unitary space, and therefore can be used to integrate functions over all of $\mathsf{U}(d)$.

To better understand the role that the uniform Haar measure plays in twirling, consider the simple example of the integral of some function $f(r, \theta, \phi)$ in spherical coordinates over all of $\mathbb{R}^3$,

$$V = \iiint_{\mathbb{R}^3} f(r, \theta, \phi) r^2 \sin(\theta) dr d\theta d\phi \,. \tag{C1}$$

Here, $r^2 \sin(\theta) dr d\theta d\phi$ is the Lebesgue *measure* over $\mathbb{R}^3$, which ensures that the integral is taken uniformly over all of $\mathbb{R}^3$. For the special unitary group in 2 dimensions, $\mathsf{SU}(2)$—the relevant group for single-qubit gates—the Haar measure is given as

$$d\mu = \sin(\theta) d\theta d\phi d\gamma \,, \tag{C2}$$

which is nearly identical to the Haar measure for a sphere, except it contains no radial component and instead includes an additional phase term γ, which comes from the U_3 parametrization of single-qubit rotations [see Eq. (352)]. While the exact Haar measure in $\mathsf{SU}(d)$ is needed for integrating over the entire unitary space that is relevant to qudits, some knowledge of the Haar measure is sufficient for the task of twirling, which only requires that we sample enough points uniformly at random that *approximate* $\mathsf{SU}(d)$.

2. Unitary *t*-designs

Twirling involves averaging a channel $\mathcal{E}$ over a unitary group $\mathsf{U}(d)$. However, any continuous group has an infinite

number of elements. For example, in the case of SU(2), there are an infinite number of points on the surface of the Bloch sphere. Therefore, it is impossible to average $\mathcal{E}$ over *all* of U(d). Instead, one typically samples unitaries from some subgroup $\mathbb{G} \subset \mathsf{U}(d)$, which *approximates* U(d). This forms the basis of what is called a *unitary t-design*, which describes a unitary group that simulates the statistical properties of uniformly distributed Haar random matrices [427] up to the t'th moment.

Classically, the notion of spherical t-designs defines a finite collection of points on the surface of a unit sphere that provide a "good" approximation to the integral over the entire unit sphere [428]. *Unitary* t-designs are the extension of spherical t-designs to the quantum domain, for which we desire to reproduce the basic properties of an entire unitary group U(d). Formally, a unitary t-design in d dimensions is a set of unitary operators $\{U_1, \ldots, U_K\}$ such that the sum over every polynomial $P_{t,t}(U_k) = U_k^{\otimes t} \otimes (U_k^*)^{\otimes t}$ of degree no larger than t in the matrix elements of U and their complex conjugates is equal to the integral of $P_{(t,t)}(U)$ over U(d),

$$\frac{1}{K}\sum_{k=1}^{K} P_{(t,t)}(U_k) = \int_{\mathsf{U}(d)} d\mu(U) P_{(t,t)}(U)\,. \tag{C3}$$

Very heuristically, unitaries in a t-design are evenly spaced around the d-dimensional unit sphere defining U(d), with larger values of t defining more densely spaced points. For example, the d-dimensional Pauli group forms a unitary 1-design, and the d-dimensional Clifford group ($\mathbb{C}_{D,n}$) forms a unitary 2-design when the qudit dimension D is prime [429]. According to Ref. [430], a set of unitaries $\{U_k\}_{k=1}^K$ forms a unitary 2-design if and only if

$$\frac{1}{K^2}\sum_{k,k'=1}^{K} \left|\mathrm{Tr}(U_{k'}^\dagger U_k)\right|^4 = 2\,. \tag{C4}$$

3. Twirling quantum channels

To understand how one constructs an average quantum channel by twirling, first consider a quantum channel $\mathcal{E}$ that represents (the error in) some quantum gate or process. Next, consider a unitary operator $\hat{U}$ that belongs to U(d). Suppose that $\mathcal{E}$ is conjugated by $\hat{U}$, mapping $\mathcal{E} \mapsto \hat{U} \circ \mathcal{E} \circ \hat{U}^\dagger$ (see Fig. 51). Using this notation, a twirled channel $\bar{\mathcal{E}}$ is given by

$$\bar{\mathcal{E}} = \int_{\mathsf{U}(d)} d\mu(\hat{U}) \hat{U} \circ \mathcal{E} \circ \hat{U}^\dagger\,. \tag{C5}$$

Thus, $\bar{\mathcal{E}}$ can be thought of as the expected value of $\mathcal{E}$ conjugated with all possible unitaries $\hat{U} \in \mathsf{U}(d)$. Because $\mathcal{E} = \mathcal{E}(\rho)$ is a linear map on a quantum state ρ, $\hat{U}$ also acts

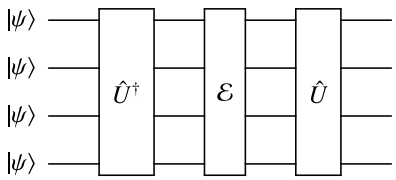


FIG. 51. Twirling. Twirling a quantum channel results in the map $\mathcal{E} \mapsto \hat{U} \circ \mathcal{E} \circ \hat{U}^\dagger$.

on ρ by conjugation:

$$\hat{U}(\rho) = U\rho U^\dagger\,, \quad \hat{U}^\dagger(\rho) = U^\dagger \rho U\,. \tag{C6}$$

Therefore, the twirled channel of a density operator $\bar{\mathcal{E}}(\rho)$ can be written

$$\bar{\mathcal{E}}(\rho) = \int_{\mathsf{U}(d)} d\mu(U) U^\dagger \mathcal{E}(U\rho U^\dagger) U\,. \tag{C7}$$

As discussed in the previous section, it is not practical to twirl over all of a unitary group U(d). Rather, it is much more common to twirl a channel over a discrete set of unitaries that approximates some properties of U(d). For example, consider the channel $\mathcal{E}(\rho) = A\rho B$, where $\{A, B\}$ are arbitrary linear operators. Next, consider some group $\{U_k\}_{k=1}^K$ consisting of K unitary operators. In the discrete case, the twirled channel $\bar{\mathcal{E}}(\rho)$ can be written as the weighted average over all K operators [282]:

$$\bar{\mathcal{E}}(\rho) = \frac{1}{K}\sum_{k=1}^{K} U_k^\dagger A U_k \rho U_k^\dagger B U_k\,. \tag{C8}$$

Note that twirling a channel does not change the average gate fidelity or process fidelity of the channel. To see this, we replace $\mathcal{E}(\rho)$ in Eq. (220) with the twirled channel $\bar{\mathcal{E}}(\rho)$ in Eq. (C7), and find that

$$F_{\mathrm{avg}}(\bar{\mathcal{E}}) = \int d\psi \int d\mu(U) \langle\psi| U^\dagger \mathcal{E}(U\rho U^\dagger) U |\psi\rangle\,, \tag{C9}$$

$$= \int d\mu(U) \int d\psi \langle\psi| U^\dagger \mathcal{E}(U\rho U^\dagger) U |\psi\rangle\,, \tag{C10}$$

$$= \int d\mu(U) F_{\mathrm{avg}}(\mathcal{E})\,, \tag{C11}$$

$$= F_{\mathrm{avg}}(\mathcal{E})\,, \tag{C12}$$

where, in the second to last step we made a change of variables $|\psi'\rangle \equiv U|\psi\rangle$, and in the final step utilized the fact that $\int d\mu(U) = 1$.

4. Pauli twirling

One can twirl a channel $\mathcal{E}$ over any group. However, it is often convenient to choose a particular group, such as the Pauli or Clifford group (see Appendix B). Because we often represent our channels in the Pauli basis (e.g., in the PTM representation; see Sec. II C 3), it is educational to first understand the basics of Pauli twirling, before considering twirling over any other unitary group. Pauli twirling an arbitrary channel $\mathcal{E}$ can be understood with the following example (shown in Fig. 53): consider the PTM of a $R_x(\pi/18)$ rotation [Fig. 53(d)], which contains off-diagonal terms only in the lower right-hand block of the PTM. When conjugating $R_x(\pi/18)$ with a Pauli from the Pauli group $P \in \{I, X, Y, Z\}$, the off-diagonal elements of $R_x(\pi/18)$ remain unchanged for $I \circ R_x(\pi/18) \circ I$ and $X \circ R_x(\pi/18) \circ X$, but have their signs flipped for $Y \circ R_x(\pi/18) \circ Y$ and $Z \circ R_x(\pi/18) \circ Z$. More generally, for any arbitrary channel $\mathcal{E}$, the signs of the off-diagonal terms remain the same for the elements of $\mathcal{E}$ with which P commutes, and are reversed for the elements of $\mathcal{E}$ with which P anticommutes.

When twirling with respect to the uniform distribution over the Pauli group, using Eq. (C8) we may write the twirled channel as

$$\bar{\mathcal{E}}(\rho) = \frac{1}{d^2} \sum_{P \in \mathbb{P}_n} P^\dagger \mathcal{E}(P\rho P^\dagger) P \,. \tag{C13}$$

However, it is often unnecessary (and inefficient) to twirl over the entire Pauli group, depending on the size of the system we are considering. In general, Pauli twirling is implemented by averaging over N randomly sampled Paulis,

$$\bar{\mathcal{E}}(\rho) = \frac{1}{N} \sum_{P \in_R \mathbb{P}_n}^{N} P^\dagger \mathcal{E}(P\rho P^\dagger) P \,, \tag{C14}$$

where R denotes that P is chosen at random from the n-qubit Pauli group $\mathbb{P}_n$ each time. In this case, the off-diagonal terms of $\mathcal{E}(\rho)$ change sign with a 50% probability upon conjugation with a randomly selected Pauli. When averaging a channel over N Paulis, the magnitudes of the off-diagonal terms scale as $\theta/\sqrt{N}$, reminiscent of a random walk, and thus vanish as $N \longrightarrow \infty$ or if by luck the correct Paulis were sampled that average to zero [see Fig. 53(h)]. This feature of twirling is often referred to as "noise tailoring," as the noise profile of a channel is modified as the number of averages increases. In fact, in the limit of $N \longrightarrow \infty$, any arbitrary Markovian error channel is mapped into a stochastic Pauli channel (see Sec. III E) by Pauli twirling. As a concrete example, in Fig. 52 we plot a random CPTP two-qubit PTM and show how the off-diagonal terms are averaged to zero under Pauli twirling as N is increased from 10, to 100, to 10^4 (see Fig. 54). Note

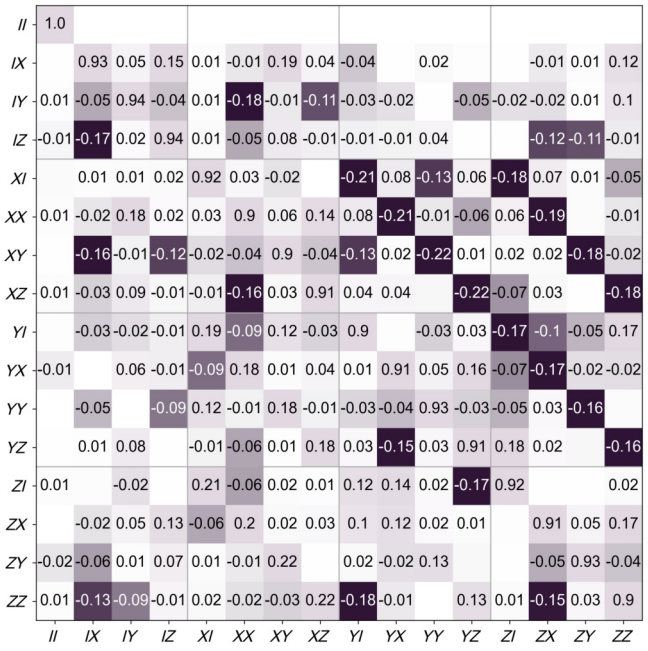


FIG. 52. PTM of a random two-qubit CPTP channel. The color (transparency) of each cell is determined by the sign (magnitude) of each entry.

that the diagonal terms in the PTM remain unchanged for all N. More specifically, Pauli twirling tailors all noise into Pauli channels, in which the diagonal entries of the PTM remain unchanged [Eq. (170)].

5. Clifford twirling

The n-qubit Clifford group $\mathbb{C}_n$ normalizes the n-qubit Pauli group. Functionally, this means that Clifford gates $C \in \mathbb{C}_n$ map Paulis to Paulis under conjugation: $CPC^\dagger \mapsto P' \in \mathbb{P}_n$, $\forall P \in \mathbb{P}_n$; see Sec. B 3 for several examples. Twirling over the entire n-qubit Clifford group can be done discretely for any channel $\mathcal{E}$:

$$\bar{\mathcal{E}}(\rho) = \frac{1}{|\mathbb{C}_n|} \sum_{C \in \mathbb{C}_n} C^\dagger \mathcal{E}(C\rho C^\dagger) C \,. \tag{C15}$$

Because the Clifford group forms a unitary 2-design, Clifford twirling replicates the properties of twirling over all of $\mathsf{U}(d)$ up to the second moment. In practice, however, one does not twirl over the entire n-qubit Clifford group due to the large number of Clifford elements (see Appendix B 3). Similar to Pauli twirling, sampling N Clifford gates at random is usually sufficient for suitably large N:

$$\bar{\mathcal{E}}(\rho) = \frac{1}{N} \sum_{C \in_R \mathbb{C}_n}^{N} C^\dagger \mathcal{E}(C\rho C^\dagger) C \,, \tag{C16}$$

where R denotes that each C is selected uniformly at random from $\mathbb{C}_n$.

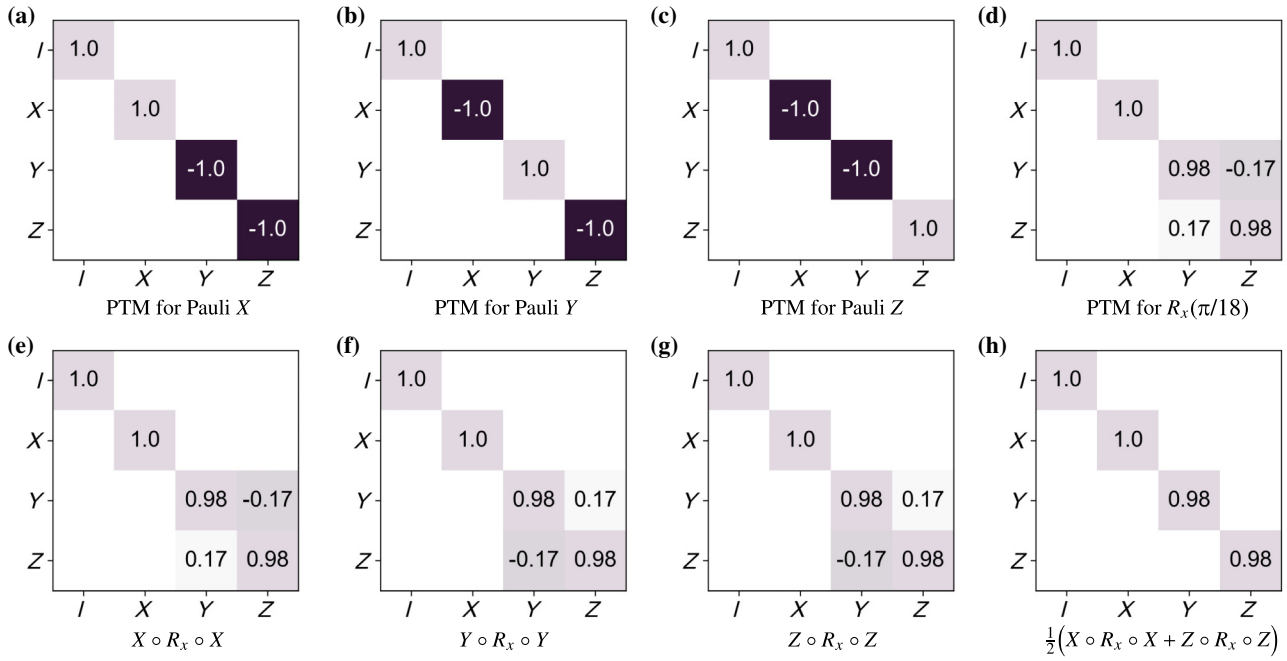


FIG. 53. Basics of Pauli twirling. (a) PTM for the Pauli-X gate. (b) PTM for the Pauli-Y gate. (c) PTM for the Pauli-Z gate. (d) PTM for an $R_x(\pi/18)$ rotation. (e) $R_x(\pi/18)$ conjugated with X gates; because $R_x(\pi/18)$ commutes with X, the PTM is left unchanged. (f) $R_x(\pi/18)$ conjugated with Y gates; because $R_x(\pi/18)$ does not commutes with Y, the signs of the off-diagonal terms have been flipped relative to $R_x(\pi/18)$. (g) $R_x(\pi/18)$ conjugated with Z gates; because $R_x(10^\circ)$ does not commutes with Z, the signs of the off-diagonal terms have been flipped relative to $R_x(\pi/18)$. (h) Average of $R_x(\pi/18)$ twirled with the Pauli-X and and Pauli-Z gates, resulting in a PTM in which the off-diagonal terms have been exactly averaged to zero. For all plots, the color (transparency) of each cell is determined by the sign (magnitude) of each entry.

Clifford twirling a quantum channel $\mathcal{E}$ has the same effect as Pauli twirling on the off-diagonal matrix elements of $\mathcal{E}$. Namely, in the limit of large N, all off-diagonal elements are averaged to zero. This is demonstrated in Fig. 54, where we show the impact of Clifford twirling on the random PTM shown in Fig. 52 for $N = 10, 100, 10^4$. However, we observe that in contrast to Pauli twirling, Clifford twirling does not preserve the eigenvalues of the PTM. Rather, Clifford twirling averages all diagonal elements of the PTM to the same value (except for the first element), effectively tailoring noise into a global depolarizing channel (see Secs. III D and IV C 4). This is due to the fact that conjugating a Pauli operator by Clifford gates can map the Pauli into a different Pauli. Thus, in the limit of large N, Clifford twirling averages the eigenvalues of the PTM. Note that, similar to Pauli twirling, Clifford twirling does not change the process fidelity of a PTM.

6. Weyl twirling

The Weyl-Heisenberg group forms a unitary 1-design. Therefore, it is possible to use Weyl operators to twirl in higher dimensions, tailoring noise into stochastic Weyl channels,

$$\mathcal{E}(\rho) = \sum_{\bar{p},\bar{q}}^{D^{2n}} \Pr(W_{\bar{p},\bar{q}}) W_{\bar{p},\bar{q}} \rho W_{\bar{p},\bar{q}}^{\dagger}, \tag{C17}$$

where ρ is an n-qudit state, $W_{\bar{p},\bar{q}} = \otimes_{i=1}^{n} W_{q_{k_i},p_{k_i}}$ is a tensor product of single-qudit operators in the D-dimensional Weyl-Heisenberg group, and $\Pr(W_{\bar{p},\bar{q}})$ is the probability of the Weyl error $W_{\bar{p},\bar{q}}$ occurring. An important note about twirling in higher dimensions is that it is just as efficient as twirling in $D = 2$, as shown in Fig. 55, where we demonstrate the numerical results of twirling away off-diagonal elements (e.g., coherent errors) in qudit transfer matrices using Weyl twirling for different qudit dimensions. In fact, this is guaranteed by Hoeffding's inequality [432], and in the context of QCVV it means that no additional sampling is required relative to qubit-based methods to achieve the same degree of noise tailoring [431].

APPENDIX D: QCVV FOR QUDITS

While the focus of this tutorial has been on characterizing and benchmarking the performance of qubit-based quantum computers, in recent years there have been significant efforts in realizing qudit-based quantum processors

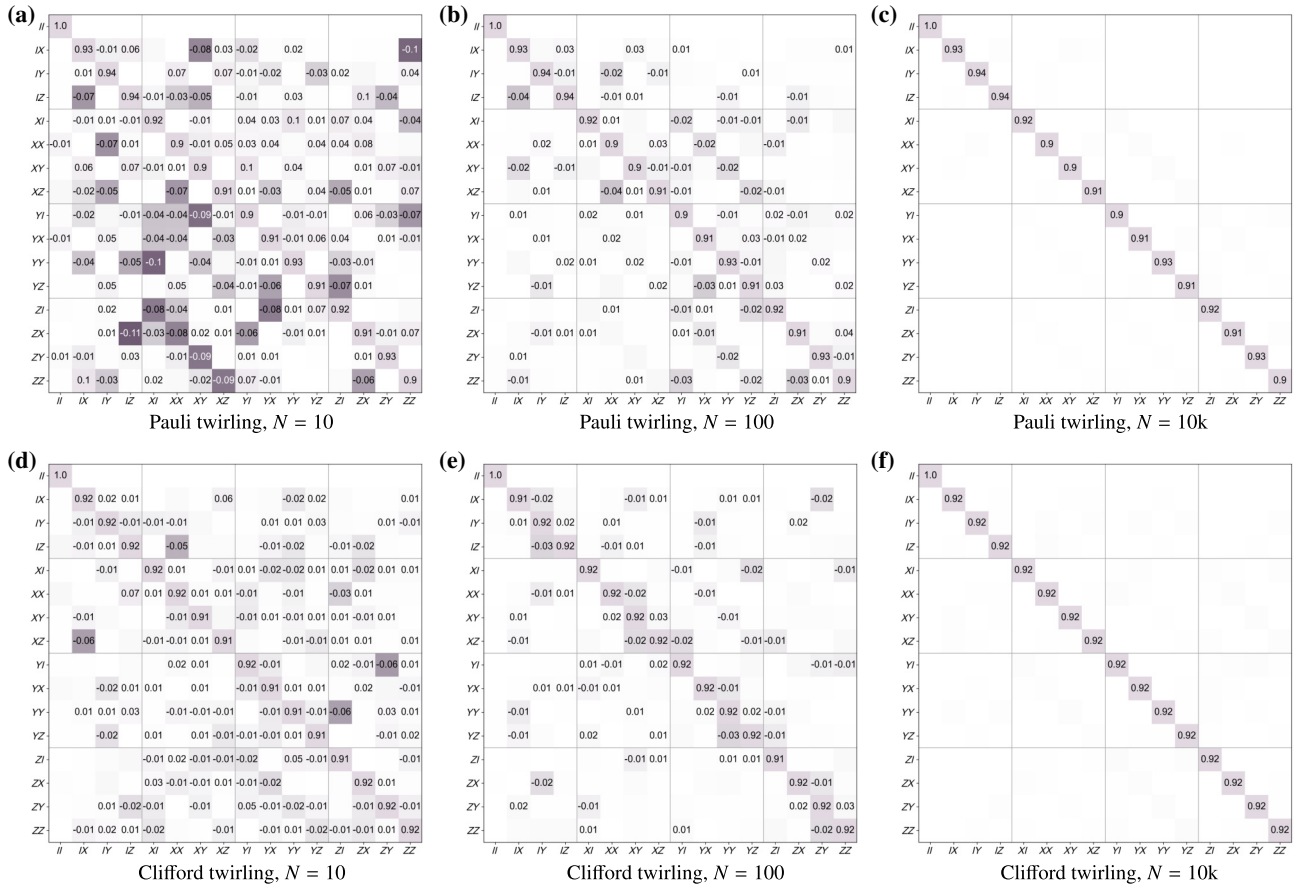


FIG. 54. Pauli twirling vs. Clifford twirling. Twirling of the random CPTP channel in Fig. 52 as a function of the number N of randomly selected Pauli and Clifford gates. Pauli twirling preserves the eigenvalues along the diagonal of the PTM, whereas Clifford twirling averages them together into a global depolarizing channel.

on platforms including superconducting circuits [8,433–435], trapped ions [436,437], and photonic circuits [438, 439]. Building a quantum computer based on qudits can yield significant advantages, such as improved quantum error correction [440–443], more efficient quantum algorithms [444], and naturally tailored quantum simulations [8,445]. To benchmark a qudit-based quantum computer, it is first incumbent upon us to generalize much of the machinery that has already been developed for qubits. In Sec. B 4, we generalized the qubit Pauli and Clifford operators for qudits. In this section, we introduce qudit transfer matrices, and then discuss tomographic reconstruction and randomized benchmarks generalized for qudits.

1. Qudit transfer matrices

Having constructed suitable bases for describing the unitary operations of qudits in Sec. B 4, we can turn our attention to generalizing transfer matrices for qudits as well (see Sec. II C for more information). Unsurprisingly, the requirements for quantum operations describing real, physical processes do not change for qudits, i.e., they must be CPTP maps. First, we can expand a qudit density matrix ρ in terms of the n-qudit Gell-Mann basis $\mathbb{G}_{D,n}$,

$$\rho = \sum_{G\in\mathbb{G}_{D,n}} \rho_G G \tag{D1}$$

where $\rho_G = \langle\langle G|\rho\rangle\rangle/D^n$ are the expansion coefficients, which can be *vectorized* into a $D^{2n}\times 1$ column vector. Now, any map $|\rho'\rangle\rangle = \Lambda\,|\rho\rangle\rangle$ can be completely described by a $D^{2n}\times D^{2n}$ transfer matrix with elements

$$\Lambda_{ij} = \frac{1}{D^n}\mathrm{Tr}[G_i\mathcal{E}(G_j)]\,, \tag{D2}$$

where $\mathcal{E}$ denotes the Kraus map defined by Eq. (63).

2. Qudit tomography

a. Qudit state tomography

Similar to qubits, the tomographic reconstruction of a qudit density matrix ρ requires an informationally complete set of qudit basis measurements (e.g., Gell-Mann or Weyl-Heisenberg bases). This requires $D^{D^n}-1$ independent experiments, from which we can reconstruct the

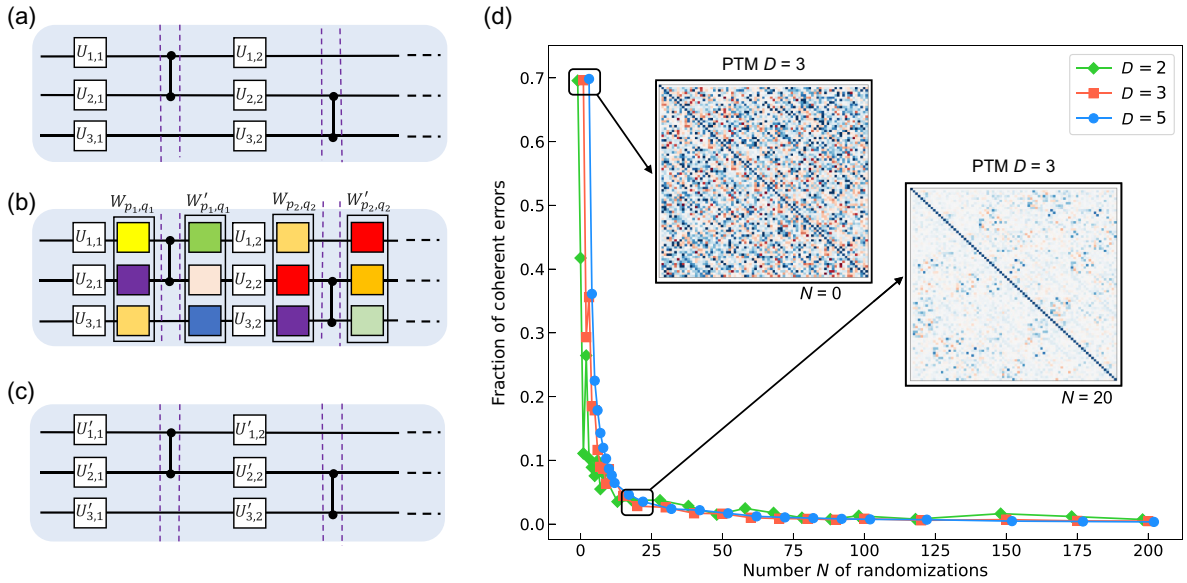


FIG. 55. Weyl twirling. (a) A hypothetical input circuit, which alternates cycles of one- and multiqudit gates. (b) Randomized compiling of the circuit in (a). Random Weyl gates ($W_{\vec{p}_j,\vec{q}_j}$) and their inverses ($W'_{\vec{p}_j,\vec{q}_j} = W^\dagger_{\vec{p}_j,\vec{q}_j} H_j W_{\vec{p}_j,\vec{q}_j}$) are added before and after every cycle of multiqudit gates H_j, respectively, to generate logically equivalent circuits. (c) Before executing the circuit, the twirling gates are recompiled into the existing native one-qudit gates. In this way, the returned circuit has the same depth as the input one. (d) Numerical study of the fraction of coherent errors under Weyl twirling in randomly generated two-qudit (Weyl) transfer matrices for $D \in \{2,3,5\}$. All transfer matrices are generated with a coherent fraction of 70%. The numerics demonstrate that twirling has the same overhead regardless of qudit dimension: suppression of off-diagonal terms in the transfer matrices for all dimensions D is approximately equal as a function of the number of random of randomizations N. The inset transfer matrices visualize the suppression of off-diagonal terms as a function of N for $D = 3$. (Figure reproduced with permission from [431].)

density matrix:

$$\rho = \frac{1}{D^n} \sum_{G \in \mathbb{G}} \langle G \rangle \, G \,. \tag{D3}$$

From these measurements, as described in Sec. VII A, the density matrix ρ can be estimated using MLE (see Sec. VII A 1). In practice, the single-qudit operations necessary to reconstruct an arbitrary qudit density matrix can be considered as the local projections onto all the computational states $|0\rangle, \ldots, |D\rangle$, as well as the $\sqrt{X^{jk}}, \sqrt{Y^{jk}}$ projections over all local two-level subspaces of the qudit. The results of experimentally reconstructed qudit Bell states ($|\psi\rangle = \frac{1}{\sqrt{D}} \sum_{i=0}^{D-1} |ii\rangle$) for qudit dimension $D = 2, 3, 4$ can be seen in Fig. 56.

b. Qudit process tomography

As with state tomography, quantum process tomography (QPT) can also be readily generalized to describe how a qudit operation maps input qudit states to output qudit states. Following the procedure described in Sec. VII B, we can use the set of informationally complete operations outlined in Sec. D 2 a for both the state preparations and measurement bases to tomographically reconstruct the qudit transfer matrix Λ. For example, Fig. 57 shows the results of performing QPT on a two-qutrit $\mathrm{CZ}^\dagger$ gate in the Gell-Mann basis [433], where $U_{\mathrm{CZ}^\dagger} = \sum_{i,j \in \mathbb{Z}_3} \omega^{-ij} \, |i,j\rangle\langle i,j\,|$. This required 81 different two-qutrit input states prepared using the following set of native gates on each qutrit: $\{I, \sqrt{X^{01}}, \sqrt{Y^{01}}, X^{01}, X^{12}X^{01}, Y^{12}\sqrt{X^{01}}, \sqrt{X^{12}X^{01}}, \sqrt{Y^{12}X^{01}}, X^{12}\sqrt{X^{01}}\}$. The same set of native gates is then used to perform state tomography on each input state, and the qudit transfer matrix Λ is reconstructed using MLE. In Fig. 57, the tomographically reconstructed transfer matrix Λ_{exp} and the error matrix $\Lambda^\dagger_{\mathrm{ideal}}\Lambda_{\mathrm{exp}}$ are displayed in the qutrit Gell-Mann basis. From these results, the process fidelity is calculated as $F_e = \mathrm{Tr}[\Lambda^\dagger_{\mathrm{ideal}}\Lambda_{\mathrm{exp}}]/d^2 = 93.2\%$. We note that the discrepancy between the process fidelity of the $\mathrm{CZ}^\dagger$ calculated via QPT and randomized benchmarks (introduced in the following section) can be attributed to SPAM errors.

c. Qudit gate set tomography

Finally, we note that gate set tomography (GST) (see Sec. VII D) can also be generalized for qudits, and has also been applied to study single-qutrit gates in Ref. [447], where it demonstrated good agreement between other SPAM-free characterization methods such as qutrit randomized benchmarking. Additionally, qudit-based GST methods can provide insight into non-Markovian errors

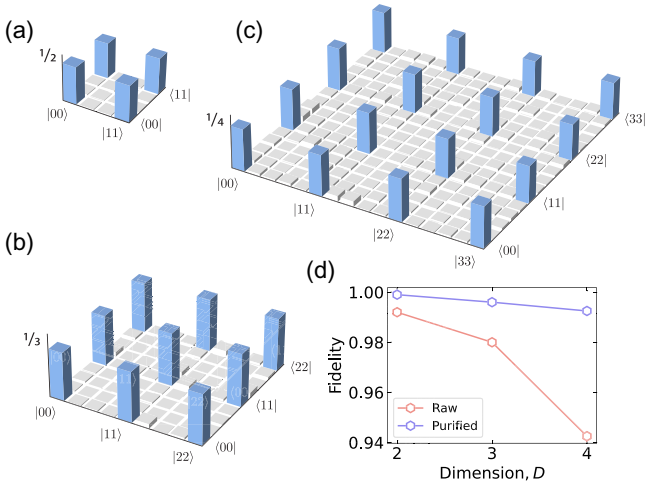


FIG. 56. Qudit state tomography. The real component of experimentally reconstructed density matrices for a (a) two-qubit Bell state $|\psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, (b) two-qutrit Bell state $|\psi\rangle = \frac{1}{\sqrt{3}}(|00\rangle + |11\rangle + |22\rangle)$, and (c) two-ququart Bell state $|\psi\rangle = \frac{1}{\sqrt{4}}(|00\rangle + |11\rangle + |22\rangle + |33\rangle)$. (d) The state fidelities for the raw and purified [446] density matrices. (Figure reproduced with permission from Ref. [272].)

and fine-grained error budgets for qudit gates, which are difficult to extract from lighter-weight methods such as randomized benchmarks.

3. Qudit randomized benchmarks

As with qubit-based randomized benchmarks (see Sec. VIII), qudit-based randomized benchmarking techniques employ random circuits to enable the efficient characterization of quantum gate sets. Broadly speaking, all of these methods leverage twirling (see Sec. C 6) to tailor noise into Pauli channels or a global depolarizing channel, yielding efficient estimates of process fidelities. In this section, we describe the generalizations required for performing randomized benchmarking, cycle benchmarking, and cross-entropy benchmarking on a qudit-based quantum computer.

a. Qudit randomized benchmarking

Having already defined the qudit Weyl-Heisenberg group and Clifford group in Sec. B 4, it is now possible to describe qudit randomized benchmarking (RB; see Sec. VIII for a background on qubit RB). The procedure for qudit randomized benchmarking follows exactly as in the qubit case, where now the random Clifford gates are sampled uniformly from $\mathbb{C}_{D,n}$, with the final gate in any RB sequence chosen to decompose the entire circuit to the identity (or up to a random Weyl operator). In general, the ground state is prepared, and the success probability is fit to an exponential decay of the qudit Z expectation value, calculated as

$$\langle Z\rangle_D = \sum_{i=0}^{D-1} p(|i\rangle)\omega^i, \tag{D4}$$

where ω is again the Dth root of unity. The average Clifford gate fidelity is then calculated from fitting the exponential $\langle Z\rangle_D(m) = Af^m$ by measuring different circuit depths

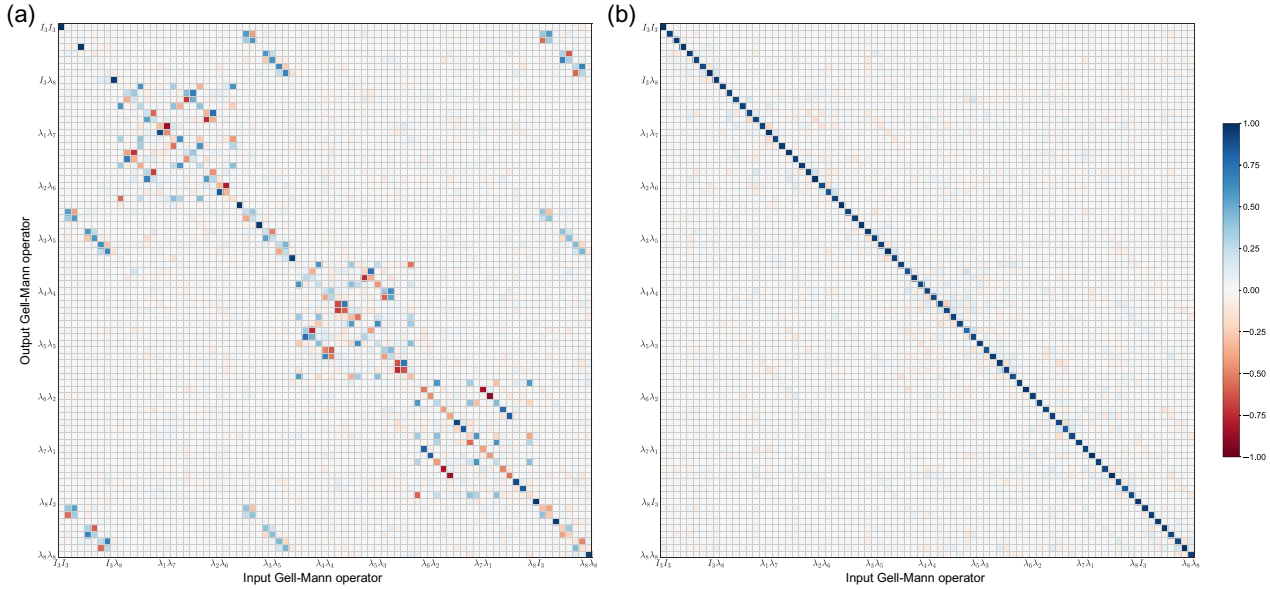


FIG. 57. Qudit process tomography. The transfer matrix of an experimentally realized two-qutrit $CZ^\dagger$ gate [433]. (a) The transfer matrix (Λ_{exp}) in the qutrit Gell-Mann basis. (b) The error matrix $\Lambda_{\text{exp}}^\dagger \Lambda_{\text{ideal}}$, with corresponding process fidelity of 93.2%. (Figure reproduced with permission from [433].)

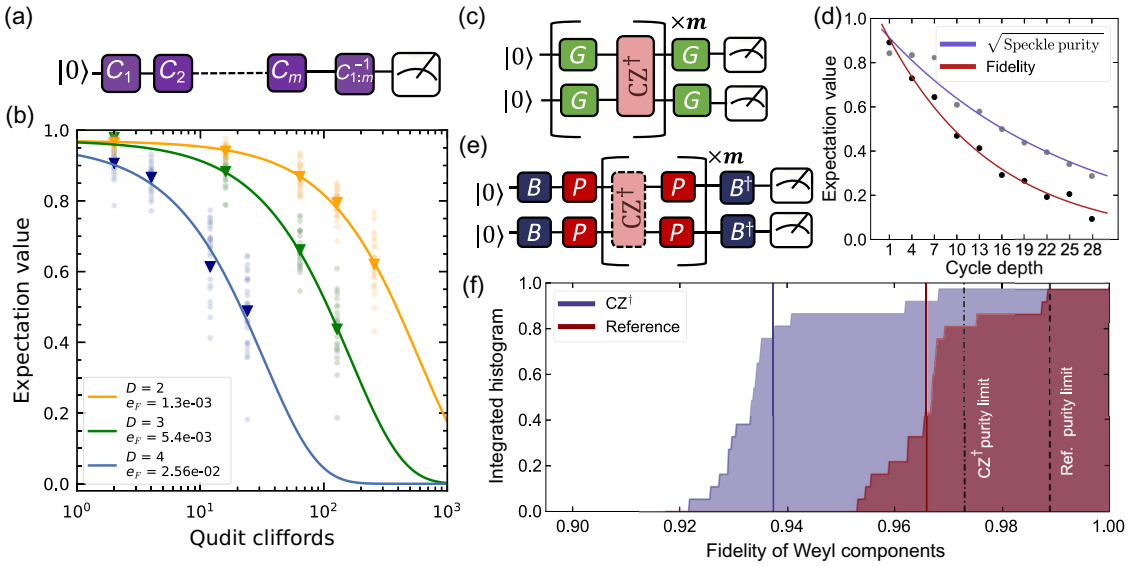


FIG. 58. Qudit randomized benchmarks. (a) Single-qudit Clifford RB circuits where m different Clifford gates C_i (purple) are selected from $\mathbb{C}_D$, after which the entire circuit is inverted with a single additional Clifford $C_{1:m}^{-1}$. (b) The experimental results of qudit RB on a superconducting qudit for $D = \{2, 3, 4\}$ [272]. The process infidelity e_F is listed in the legend for each dimension D. (c) Circuit schematic of XEB. A CZ† gate is interleaved between m cycles of random SU(3) gates (green). (d) A dressed CZ† fidelity of 0.933(3) was measured from the exponential decay of the XEB results [433]. Additionally, the speckle-purity limited fidelity of the dressed cycle was estimated to be 0.961(3). (e) Circuit schematic of CB. The system is prepared in a Weyl basis state B (blue), after which the CZ† gate is interleaved between m cycles of random Weyl gates (red). Finally, the system is rotated back to the original Weyl basis with an additional final cycle of inverting Weyl gates $B^\dagger$. (f) An integrated histogram of CB for both the CZ† gate and a reference cycle from Ref. [433], with the solid vertical lines giving the process fidelities of 0.936(1) and 0.966(1), respectively. Together, these yield an estimated gate fidelity of 97.3(1)% for the CZ† gate. Moreover, one can extract an error budget directly from CB, giving purity-limited fidelities of 0.973(9) and 0.989 (with negligible error) for the dressed CZ† and reference cycles, respectively; together, this gives an estimated purity limit 0.986(9) for the isolated CZ† gate. (Figures reproduced with permission from [272,433].)

m and converting the process polarization f to an average gate fidelity or process fidelity (see Table II). We note that although the qudit Z operator is in general non-Hermitian, its phase does not change under depolarization. Therefore, the imaginary component—which is initially zero due to preparing in the ground state—remains zero throughout.

Initial demonstrations of qudit randomized benchmarking have been performed for $D = 3$ in Refs. [298,431] and $D = 4$ in Refs. [434,448]. In Figs. 58(a) and 58(b), the circuits and results of performing RB on the same qudit operating in $D = \{2, 3, 4\}$ is shown [272], yielding average Clifford process fidelities of $F_e = \{0.99872(1), 0.9946(2), \text{ and } 0.974(2)\}$. Since there are in general $\{2, 6, 12\}$ native gates (excluding software defined virtual Z gates) needed to decompose Clifford gates in $D = \{2, 3, 4\}$, from the Clifford fidelities one can calculate the average native gate process fidelities, yielding $F_e = \{0.99936(3), 0.99909(4), 0.9978(2)\}$ for the results in Fig. 58. We further note that interleaved RB and simultaneous RB can also be performed to obtain additional insight into specific qudit gate performance as well as effects from undesired crosstalk interactions [298].

b. Qudit cycle benchmarking

Cycle benchmarking (CB) is useful for qudit-based QCVV, specifically in the context of multiqudit gates, as performing randomized benchmarking on a multiqudit gate requires sampling and decomposing multiqudit Clifford gates, often requiring many native multiqudit gates. In contrast, CB can be performed with significantly fewer multiqudit gates per circuit and can be scaled to studying larger systems (see Sec. VIII G). For qudit CB, the eigenstates and twirling gates are chosen from the Weyl-Heisenberg group. Notably, each $W \in \mathbb{W}_\mathbb{D}$ commutes with its Hermitian conjugate $W^\dagger = W^{D-1}$, which also belongs to the Weyl-Heisenberg group. This implies that these operators share the same eigenbasis, such that $\langle W^\dagger \rangle = \langle \bar{W} \rangle$, which allow us to reduce the number of required measurements needed to characterize the Weyl decays [298].

Qutrit CB was first described and demonstrated in Ref. [298], and was later used to benchmark two-qutrit CZ and CZ† gates with interleaved gate fidelities as high as 95.2(3)% and 97.3(1)%, respectively [433]. The circuits and integrated histogram results of the CB protocol for a CZ† gate are shown in Figs. 58(e) and 58(f).

c. Qudit cross-entropy benchmarking

The cross-entropy benchmarking (XEB) protocol (see Sec. VIII C 4) can likewise be straightforwardly generalized to qudits. In the case of qudit XEB, the local twirling is now performed via Haar-random $\mathsf{SU}(d)$ gates, and the analysis is performed on the output ditstring (rather than bitstring) results. Qutrit XEB has been described and performed in Ref. [433], where it was used to benchmark a two-qutrit $\mathrm{CZ}^\dagger$ gate. Those results can also be found in Figs. 58(c) and 58(d), where we also provide a circuit schematic for XEB sequences. We note here that the dressed gate fidelities for the $\mathrm{CZ}^\dagger$ estimated from qutrit CB and XEB agree to within error bars, which is expected due to the fact that twirling does not change the average gate fidelity or process fidelity of a gate (see Sec. C 3).

APPENDIX E: GAUGE AMBIGUITY IN PAULI NOISE LEARNING

One prominent advantage of cycle benchmarking (CB; see Sec. VIII G) or cycle error reconstruction (CER; see Sec. IX C 1) is the intrinsic robustness to SPAM errors, which is a general feature of randomized-benchmarking-like protocols. However, when benchmarking cycles containing multiqubit Clifford gates, CER generally cannot resolve every Pauli fidelity (or Pauli error rate) individually. Instead, for a certain subset of Pauli operators, only the geometric mean of Pauli fidelities can be estimated. See Fig. 39 for an example. This issue of "degeneracy" is not a drawback of any specific method, but is related to the fundamental notion of gauge ambiguity [156] (see Sec. II E 2). That is, when taking unknown SPAM noise into account, there exist certain gauge degrees of freedom in the noise model that cannot be resolved. In this section, we discuss a theory [399] that fully characterizes the gauge-consistently learnable information in Pauli noise learning.

We start with four assumptions about the noise: (1) any single-qubit unitary gate can be implemented perfectly; (2) a set of multiqubit Clifford gates $\{G\}$ can be implemented with gate-dependent Pauli noise, i.e., $\widetilde{G} = G \circ \Lambda_G$, where $\widetilde{G}$ denotes the noisy gate and Λ_G the transfer matrix capturing the Pauli noise; (3) any state preparation and POVM measurement can be implemented subject to an unknown Pauli noise channel; (4) the Pauli fidelities of all Pauli channels are strictly positive. For the first condition, we can allow single-qubit gate cycles to have gate-independent noise, which are standard assumptions of CB and CER, but in that case one can simply absorb the noise into the multiqubit Clifford gate (known as dressed cycles [398]). The second and third assumptions can be guaranteed via randomized compiling [154,155]. The last one is for regularization and should hold for any reasonable gate set. Now, we ask the following question: what information of $\{\Lambda_G\}$ can be learned in a SPAM-robust manner despite the existence of unknown SPAM noise?

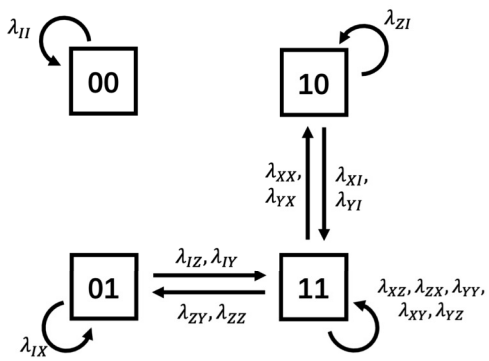
For the task of Pauli channel estimation, this question can be answered step-by-step: firstly, any individual Pauli fidelity, $\lambda_P^G \equiv \mathrm{Tr}[P\Lambda_G(P)]/2^n$, can be learned SPAM-robustly if and only if G preserves the *pattern* of P. The pattern of an n-qubit Pauli (with sign ignored) is defined as an n-bit string whose ith bit is 0 if $P_i = I$ and 1 otherwise (e.g., $XZIYI \mapsto 11010$). Take the CNOT gate as an example, which maps Paulis into Paulis (see Table VII). Since

$$\begin{aligned} \text{CNOT}: &\quad ZI \mapsto ZI\,, \quad YY \mapsto XZ\,, \quad XI \mapsto XX\,, \\ \text{pattern}: &\quad 01 \mapsto 01\,, \quad 11 \mapsto 11\,, \quad 10 \mapsto 11\,, \end{aligned}$$

we have that λ_{ZI} and λ_{YY} are SPAM-robustly learnable, while λ_{XI} is not. We can summarize how the CNOT changes the pattern of all two-qubit Pauli operators in a *pattern transform graph*, with 4 nodes and 16 edges, shown in Fig. 59. Only those Pauli fidelities which lie on a self-loop are individually learnable.

Next, a product of Pauli fidelities $(\lambda_{P_1} \cdots \lambda_{P_M})$ is SPAM-robustly learnable if the corresponding edges form a cycle in the pattern transform graph. For the example of a CNOT gate, though λ_{XI} and λ_{XX} are individually unlearnable, their product $\lambda_{XI}\lambda_{XX}$ is learnable. A more rigorous statement goes as follows: define the log-Pauli fidelities $l_P^G \equiv \log \lambda_P^G$ for all P. The space of linear functions of $\{l_P\}$ has a natural isomorphism to the edge space of the pattern transform graph. The result states that the learnable linear functions form a subspace corresponding to the cycle space of the graph, a notion from algebraic graph theory [449]. The number of learnable and unlearnable degrees of freedom can also be inferred using graph-theoretical tools. More details are presented in Refs. [399,450].

Here, we briefly sketch the proof of these results. To see that everything in the cycle space is learnable, one just needs to construct a proper CB-type experiment. Take the CNOT as an example: since the CNOT preserves ZI, to learn λ_{ZI} we first prepare an eigenstate of ZI, repeat the CNOT m times (under randomized compiling), and then measure the expectation value of ZI, from which we can fit an exponential decay of the form $A_{ZI}\lambda_{ZI}^m$, where A_{ZI} is some SPAM-dependent coefficient. Similarly, since the CNOT preserves the pattern of YY, to learn λ_{YY} we perform the same protocol; however, now we must interleave certain single-qubit gates (e.g., $\sqrt{Z} \otimes \sqrt{X}$) following each application of the CNOT, from which one can obtain $A_{YY}\lambda_{YY}^m$ (similar techniques are mentioned in [378]). Finally, to learn products like $\lambda_{XI}\lambda_{XX}$, one can simply repeat the CNOT $2m$ times and measure in the eigenstate of XX, yielding $A_{XX}(\lambda_{XI}\lambda_{XX})^m$. To see that everything outside the cycle space is unlearnable, one can show that every cut in the pattern transform graph induces a gauge transformation. For example, for a CNOT gate, a cut between 10 and the



Gate	CNOT
Basis for cycle space	$l_{II}, l_{ZI}, l_{IX}, l_{ZX}, l_{XZ}, l_{YY}, l_{XY}, l_{YZ}$, $l_{IZ}+l_{ZZ}, l_{IY}+l_{ZY}, l_{IZ}+l_{ZY}$, $l_{XI}+l_{XX}, l_{YI}+l_{YX}, l_{XI}+l_{YX}$
Learnable Pauli fidelities	$\lambda_{II}, \lambda_{ZI}, \lambda_{IX}, \lambda_{ZX}, \lambda_{XZ}, \lambda_{YY}, \lambda_{XY}, \lambda_{YZ}$, $\lambda_{IZ}\lambda_{ZZ}, \lambda_{IY}\lambda_{ZY}, \lambda_{IZ}\lambda_{ZY}$, $\lambda_{XI}\lambda_{XX}, \lambda_{YI}\lambda_{YX}, \lambda_{XI}\lambda_{YX}$
Learnable Pauli errors (up to first order)	$p_{II}, p_{ZI}, p_{IX}, p_{ZX}, p_{XZ}, p_{YY}, p_{XY}, p_{YZ}$, $p_{IZ}+p_{ZZ}, p_{IY}+p_{ZY}, p_{IZ}+p_{ZY}$, $p_{XI}+p_{XX}, p_{YI}+p_{YX}, p_{XI}+p_{YX}$
Unlearnable Pauli errors	$p_{XI}, p_{IZ}, \cdots$

FIG. 59. Learnability of Pauli noise. (Left) Pattern transform graph of CNOT. (Right) Learnable and unlearnable information of the CNOT gate.

other nodes induces the following gauge transform:

$$\lambda_{XI}, \lambda_{YI} \mapsto \kappa\lambda_{XI}, \kappa\lambda_{XI}\,,$$
$$\lambda_{XX}, \lambda_{YX} \mapsto \kappa^{-1}\lambda_{XX}, \kappa^{-1}\lambda_{YX}\,,$$

for a real number κ sufficiently close to 1. The SPAM noise needs to change correspondingly and is omitted here. One can show that this is indeed a gauge transformation that preserve all assumptions of the Pauli noise model. Since any learnable function must be orthogonal to all gauge transformations, and the cut space is orthogonal complement to the cycle space, this implies any learnable function must be inside the cycle space.

In practice, any noise channel Λ_G should be sufficiently close to identity, i.e., $\lambda_P \to 1$. In this regime, any function of Λ_G can be approximated to first order by an affine function of $\{l_b^G\}_b$, and the above result can thus be used to infer the learnability of a general function to first order, including the Pauli error rates. Interestingly, it can be shown that the first-order learnable Pauli error rates are also isomorphic to the cycle space (implicit from Ref. [398, Lemma 3] and rigorously proven in Ref. [451, Appendix D]). In other words, the cycle space of the pattern transform graph is invariant under the Walsh-Hadamard transform. As a concrete example, in Fig. 59, we list the cycle basis, learnable fidelities, first-order learnable error rates, and a possible choice of gauge parameters for the CNOT.

We end this section with some relevant observations:

(a) The unlearnablity is rooted in the gauge ambiguity of SPAM noise. If SPAM noise is very small compared to the gate noise, one can expect the ambiguity for gate noise characterization to be negligible. However, there are experiments suggesting this might not be the case for state-of-the-art quantum computing platforms [399]. Nevertheless, one can always try to bound the unlearnable parameters using physicality constraints (e.g., CPTP conditions).

(b) If we treat quantum circuits as a black box, any observable properties should, by definition, be learnable functions. Therefore, by properly characterizing all learnable degrees of freedom, one should in principle be able to perform error mitigation (see, e.g., Ref. [452]). However, such gauge-consistent error-mitigation techniques within the Pauli noise model have yet to be developed. On the other hand, there exist error-mitigation experiments based on Pauli noise learning that avoid the learnability issue by introducing additional assumptions [378,453]. It is an interesting direction to better understand the relation between noise learnability and quantum error mitigation.

(c) It is common to make additional locality assumptions about the Pauli noise channels, so that the number of noise parameters becomes manageable. Reference [450] studies how to reconcile the locality assumptions with the learnable of Pauli noise, giving a similar graph-algebraic characterization of learnable and unlearnable degrees of freedom as described here.

(d) Another practical issue for noise characterization is whether one can estimate a certain noise parameter to relative precision (using a small number of experiments). Achieving this usually requires multiple repetitions of specific gate sequence (known as "germs" in GST) to amplify the noise parameter. It has been proven that any product of Pauli fidelities corresponding to a cycle in the pattern transform graph can be learned to relative precision by repeating certain sequence of gates [398,450]. In contrast, an unlearnable function as discussed above cannot be estimated to arbitrarily small precision, no matter if it is additive precision or relative precision.

[1] P. W. Shor, in *Proceedings 35th Annual Symposium on Foundations of Computer Science* (IEEE, 1994), p. 124.

[2] F. Arute, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[3] Y. Wu, *et al.*, Strong quantum computational advantage using a superconducting quantum processor, Phys. Rev. Lett. **127**, 180501 (2021).

[4] Q. Zhu, *et al.*, Quantum computational advantage via 60-qubit 24-cycle random circuit sampling, Sci. Bull. **67**, 240 (2022).

[5] L. S. Madsen, F. Laudenbach, M. F. Askarani, F. Rortais, T. Vincent, J. F. Bulmer, F. M. Miatto, L. Neuhaus, L. G. Helt, M. J. Collins, *et al.*, Quantum computational advantage with a programmable photonic processor, Nature **606**, 75 (2022).

[6] S. Hacohen-Gourgy, L. S. Martin, E. Flurin, V. V. Ramasesh, K. B. Whaley, and I. Siddiqi, Quantum dynamics of simultaneously measured non-commuting observables, Nature **538**, 491 (2016).

[7] J. I. Colless, V. V. Ramasesh, D. Dahlen, M. S. Blok, M. E. Kimchi-Schwartz, J. R. McClean, J. Carter, W. A. de Jong, and I. Siddiqi, Computation of molecular spectra on a quantum processor with an error-resilient algorithm, Phys. Rev. X. **8**, 011021 (2018).

[8] M. S. Blok, V. V. Ramasesh, T. Schuster, K. O'Brien, J. M. Kreikebaum, D. Dahlen, A. Morvan, B. Yoshida, N. Y. Yao, and I. Siddiqi, Quantum information scrambling on a superconducting qutrit processor, Phys. Rev. X **11**, 021010 (2021).

[9] X. Mi, M. Ippoliti, C. Quintana, A. Greene, Z. Chen, J. Gross, F. Arute, K. Arya, J. Atalaya, R. Babbush, *et al.*, Time-crystalline eigenstate order on a quantum processor, Nature **601**, 531 (2022).

[10] A. Morvan, T. Andersen, X. Mi, C. Neill, A. Petukhov, K. Kechedzhi, D. Abanin, A. Michailidis, R. Acharya, F. Arute, *et al.*, Formation of robust bound states of interacting microwave photons, Nature **612**, 240 (2022).

[11] L. Xiang, W. Jiang, Z. Bao, Z. Song, S. Xu, K. Wang, J. Chen, F. Jin, X. Zhu, Z. Zhu, *et al.*, Long-lived topological time-crystalline order on a quantum processor, Nat. Commun. **15**, 8963 (2024).

[12] T. Yamakawa and M. Zhandry, in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2022), p. 69.

[13] S. Aaronson, How much structure is needed for huge quantum speedups? arXiv:2209.06930.

[14] S. Chen, J. Cotler, H.-Y. Huang, and J. Li, The complexity of NISQ, Nat. Commun. **14**, 6001 (2023).

[15] A. Anshu, N. P. Breuckmann, and C. Nirkhe, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2023), p. 1090.

[16] D. Aharonov, X. Gao, Z. Landau, Y. Liu, and U. Vazirani, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2023), p. 945.

[17] Y. Cao, J. Romero, and A. Aspuru-Guzik, Potential of quantum computing for drug discovery, IBM J. Res. Dev. **62**, 6 (2018).

[18] I. Siddiqi, Engineering high-coherence superconducting qubits, Nat. Rev. Mater. **6**, 875 (2021).

[19] A. Y. Kitaev, Quantum measurements and the Abelian stabilizer problem, arXiv:quant-ph/9511026.

[20] L. K. Grover, in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 1996), p. 212.

[21] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, SIAM Rev. **41**, 303 (1999).

[22] D. Coppersmith, An approximate Fourier transform useful in quantum factoring, arXiv:quant-ph/0201067.

[23] A. W. Harrow, A. Hassidim, and S. Lloyd, Quantum algorithm for linear systems of equations, Phys. Rev. Lett. **103**, 150502 (2009).

[24] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[25] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. **17**, 1013 (2021).

[26] A. J. Daley, I. Bloch, C. Kokail, S. Flannigan, N. Pearson, M. Troyer, and P. Zoller, Practical quantum advantage in quantum simulation, Nature **607**, 667 (2022).

[27] T. Proctor, K. Young, A. D. Baczewski, and R. Blume-Kohout, Benchmarking quantum computers, Nat. Rev. Phys. **7**, 105 (2025).

[28] The QCVV toolbox *also* includes other "tools" besides protocols. They include conceptual tools like *twirling* that are used by theorists to devise new protocols, and standardizing tools like *metrics* and *models* that enable clear communication between practitioners. But *protocols* are the heart of the field.

[29] J. Eisert, D. Hangleiter, N. Walk, I. Roth, D. Markham, R. Parekh, U. Chabaud, and E. Kashefi, Quantum certification and benchmarking, Nat. Rev. Phys. **2**, 382 (2020).

[30] M. Kliesch and I. Roth, Theory of quantum system certification, PRX Quantum **2**, 010201 (2021).

[31] In mathematics, a vector space is a Hilbert space if and only if it is isomorphic to its dual space. But *all* finite-dimensional vector spaces are Hilbert spaces, and finite-dimensional spaces suffice to describe quantum data registers. So, the mathematical implications of "Hilbert space" are an unnecessary red herring for the purposes of this tutorial.

[32] J. J. Sakurai and J. Napolitano, *Modern Quantum Mechanics* (Cambridge University Press, Cambridge, 2020), 3rd ed.

[33] This subsection intentionally presents a simplified model of quantum mechanics consistent with most undergraduate textbooks. We neglect measurements of degenerate observables, which must be modeled by projectors of rank > 1, for simplicity's sake. This important case is fully modeled by POVMs in the next subsection.

[34] This means that it is a variable in the *model* that has no physical reality, and can be varied without changing anything observable.

[35] It is "projection-valued" because it assigns a projection operator, rather than a non-negative real number, to each outcome. Born's rule, with any state $|\psi\rangle\langle\psi|$, defines a linear functional that maps a projection-valued measure to

a standard probability measure, which is the probability distribution over that measurement's outcomes.

[36] $\mathcal{B}(\mathcal{H})$ means "the space of bounded operators on $\mathcal{H}$." Sometimes $\mathcal{L}(\mathcal{H})$, meaning "the space of linear operators on $\mathcal{H}$," is used instead. These coincide when $\mathcal{H}$ is finite dimensional.

[37] Technically, a POVM is a measure (like a probability distribution) over possible events, but which is "operator valued," meaning that instead of assigning a *probability* to each event, it assigns a *positive semidefinite operator* to each event, whose inner product with the system's state ρ defines the event's probability.

[38] It is possible for initial correlation between the principal system and its environment to exist, and to be modeled. This scenario is advanced, conceptually tricky, and considered *non-Markovian*. It is not often considered in QCVV, and is outside the scope of this tutorial.

[39] W. F. Stinespring, Positive functions on c*-algebras, Proc. Am. Math. Soc. **6**, 211 (1955).

[40] C. J. Wood, J. D. Biamonte, and D. G. Cory, Tensor networks and graphical calculus for open quantum systems, Quantum Inf. Comput. **15**, 759 (2015).

[41] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, 2010).

[42] "Vectorization" is used, in quantum information science and QCVV specifically, in at least two distinct ways. This one (Eq. 65) maps a mathematical "vector" (an abstract element ρ of Hilbert-Schmidt space) to a computer programming "vector" (a concrete list of numbers). But around Eq. 111, we introduce another "vectorization" that maps mathematical vectors in $\mathcal{B}(\mathcal{H})$ to mathematical vectors in $\mathcal{H}\otimes\mathcal{H}$. Both are standard in the literature, and we apologize to the reader for our field's overloaded terminology.

[43] A. Gilchrist, D. R. Terno, and C. J. Wood, Vectorization of quantum operations and its use, arXiv:0911.2539.

[44] It is possible to define quantum operations that map $\mathcal{B}(\mathcal{H})\mapsto\mathcal{B}(\mathcal{H}')$, where $\mathcal{H}\neq\mathcal{H}'$, but these are used relatively rarely in QCVV and out of scope for this tutorial.

[45] R. Blume-Kohout, M. P. da Silva, E. Nielsen, T. Proctor, K. Rudinger, M. Sarovar, and K. Young, A taxonomy of small Markovian errors, PRX Quantum **3**, 020335 (2022).

[46] M. T. Mądzik, S. Asaad, A. Youssry, B. Joecker, K. M. Rudinger, E. Nielsen, K. C. Young, T. J. Proctor, A. D. Baczewski, A. Laucht, *et al.*, Precision tomography of a three-qubit electron-nuclear quantum processor in silicon, Nature **601**, 348 (2022).

[47] Note that some authors consider leakage, for example, to be a non-TP process, in which case the top row of the PTM captures state-dependent leakage. This is true if one only considers the qubit subspace within the full Hilbert space. However, strictly speaking, leakage is still TP, since the total probability of observing *some* outcome is preserved. For example, in some platforms leakage cannot be detected, and might instead be (erroneously) measured as 0 or 1, but the total number of shots will remain the same. In other platforms leakage can be more easily measured (see, for example, Fig. 36), in which case the total probability of observing 0, 1, or 2 is preserved. Therefore, when considering only the qubit subspace in the presence of leakage, some authors to relax the TP constraint, and instead simply require that the total probability must not increase (i.e., $\mathrm{Tr}\mathcal{E}(\rho)\leq\mathrm{Tr}\rho$).

[48] M.-D. Choi, Completely positive linear maps on complex matrices, Linear Algebra Appl. **10**, 285 (1975).

[49] E. Sudarshan, P. Mathews, and J. Rau, Stochastic dynamics of quantum-mechanical systems, Phys. Rev. **121**, 920 (1961).

[50] A. Jamiołkowski, Linear transformations which preserve trace and positive semidefiniteness of operators, Rep. Math. Phys. **3**, 275 (1972).

[51] K. Życzkowski and I. Bengtsson, On duality between quantum maps and quantum states, Open Syst. Inf. Dyn. **11**, 3 (2004).

[52] M. Hatridge, S. Shankar, M. Mirrahimi, F. Schackert, K. Geerlings, T. Brecht, K. Sliwa, B. Abdo, L. Frunzio, S. M. Girvin, *et al.*, Quantum back-action of an individual variable-strength measurement, Science **339**, 178 (2013).

[53] Y.-W. Cho, Y. Kim, Y.-H. Choi, Y.-S. Kim, S.-W. Han, S.-Y. Lee, S. Moon, and Y.-H. Kim, Emergence of the geometric phase from quantum measurement back-action, Nat. Phys. **15**, 665 (2019).

[54] E. B. Davies and J. T. Lewis, An operational approach to quantum probability, Commun. Math. Phys. **17**, 239 (1970).

[55] K. Rudinger, G. J. Ribeill, L. C. Govia, M. Ware, E. Nielsen, K. Young, T. A. Ohki, R. Blume-Kohout, and T. Proctor, Characterizing mid-circuit measurements on a superconducting qubit using gate set tomography, Phys. Rev. Appl. **17**, 014014 (2022).

[56] J. von Neumann, *Mathematische Grundlagen der Quantenmechanik* (Springer, Berlin, 1932).

[57] W. H. Zurek, Decoherence and the transition from quantum to classical, Phys. Today **44**, 36 (1991).

[58] V. B. Braginsky, Y. I. Vorontsov, and K. S. Thorne, Quantum nondemolition measurements, Science **209**, 547 (1980).

[59] V. B. Braginsky and F. Y. Khalili, Quantum nondemolition measurements: The route from toys to tools, Rev. Mod. Phys. **68**, 1 (1996).

[60] J. Volz, R. Gehr, G. Dubois, J. Estève, and J. Reichel, Measurement of the internal state of a single atom without energy exchange, Nature **475**, 210 (2011).

[61] R. Dassonneville, T. Ramos, V. Milchakov, L. Planat, É. Dumur, F. Foroughi, J. Puertas, S. Leger, K. Bharadwaj, J. Delaforce, *et al.*, Fast high-fidelity quantum nondemolition qubit readout via a nonperturbative cross-Kerr coupling, Phys. Rev. X **10**, 011045 (2020).

[62] Google Quantum AI, Suppressing quantum errors by scaling a surface code logical qubit, Nature **614**, 676 (2023).

[63] J.-G. Liu, Y.-H. Zhang, Y. Wan, and L. Wang, Variational quantum eigensolver with fewer qubits, Phys. Rev. Res. **1**, 023025 (2019).

[64] I. Cong, S. Choi, and M. D. Lukin, Quantum convolutional neural networks, Nat. Phys. **15**, 1273 (2019).

[65] C. M. Caves, K. S. Thorne, R. W. Drever, V. D. Sandberg, and M. Zimmermann, On the measurement of a weak classical force coupled to a quantum-mechanical oscillator. I. Issues of principle, Rev. Mod. Phys. **52**, 341 (1980).

[66] I. Siddiqi, R. Vijay, M. Metcalfe, E. Boaknin, L. Frunzio, R. Schoelkopf, and M. Devoret, Dispersive measurements of superconducting qubit coherence with a fast latching readout, Phys. Rev. B **73**, 054510 (2006).
[67] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, Circuit quantum electrodynamics, Rev. Mod. Phys. **93**, 025005 (2021).
[68] A. A. Clerk, M. H. Devoret, S. M. Girvin, F. Marquardt, and R. J. Schoelkopf, Introduction to quantum noise, measurement, and amplification, Rev. Mod. Phys. **82**, 1155 (2010).
[69] A. N. Korotkov, Quantum Bayesian approach to circuit QED measurement with moderate bandwidth, Phys. Rev. A **94**, 042326 (2016).
[70] K. Murch, S. Weber, C. Macklin, and I. Siddiqi, Observing single quantum trajectories of a superconducting quantum bit, Nature **502**, 211 (2013).
[71] S. Weber, A. Chantasri, J. Dressel, A. N. Jordan, K. Murch, and I. Siddiqi, Mapping the optimal route between two quantum states, Nature **511**, 570 (2014).
[72] G. Koolstra, N. Stevenson, S. Barzili, L. Burns, K. Siva, S. Greenfield, W. Livingston, A. Hashim, R. Naik, J. Kreikebaum, *et al.*, Monitoring fast superconducting qubit dynamics using a neural network, Phys. Rev. X **12**, 031017 (2022).
[73] Y. Kim, Y.-S. Kim, S.-Y. Lee, S.-W. Han, S. Moon, Y.-H. Kim, and Y.-W. Cho, Direct quantum process tomography via measuring sequential weak values of incompatible observables, Nat. Commun. **9**, 192 (2018).
[74] K. Siva, G. Koolstra, J. Steinmetz, W. P. Livingston, D. Das, L. Chen, J. M. Kreikebaum, N. Stevenson, C. Jünger, D. I. Santiago, *et al.*, Time-dependent Hamiltonian reconstruction using continuous weak measurements, PRX Quantum **4**, 040324 (2023).
[75] C. A. Fuchs and A. Peres, Quantum-state disturbance versus information gain: Uncertainty relations for quantum information, Phys. Rev. A **53**, 2038 (1996).
[76] S. Hong, Y.-S. Kim, Y.-W. Cho, J. Kim, S.-W. Lee, and H.-T. Lim, Demonstration of complete information tradeoff in quantum measurement, Phys. Rev. Lett. **128**, 050401 (2022).
[77] E. Nielsen, K. Rudinger, T. Proctor, K. Young, and R. Blume-Kohout, Efficient flexible characterization of quantum processors with nested error models, New J. Phys. **23**, 093020 (2021).
[78] K. Rudinger, C. W. Hogle, R. K. Naik, A. Hashim, D. Lobser, D. I. Santiago, M. D. Grace, E. Nielsen, T. Proctor, S. Seritan, S. M. Clark, R. Blume-Kohout, I. Siddiqi, and K. C. Young, Experimental characterization of crosstalk errors with simultaneous gate set tomography, PRX Quantum **2**, 040338 (2021).
[79] A. Hashim, S. Seritan, T. Proctor, K. Rudinger, N. Goss, R. Naik, J. M. Kreikebaum, D. Santiago, and I. Siddiqi, Benchmarking quantum logic operations relative to thresholds for fault tolerance, npj Quantum Inf. **9**, 109 (2023).
[80] R. Brieger, I. Roth, and M. Kliesch, Compressive gate set tomography, PRX Quantum **4**, 010325 (2023).
[81] O. Di Matteo, J. Gamble, C. Granade, K. Rudinger, and N. Wiebe, Operational, gauge-free quantum tomography, Quantum **4**, 364 (2020).
[82] J. P. Marceaux and K. Young, in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)* (IEEE Computer Society, Los Alamitos, CA, USA, 2023), p. 1401.
[83] J. Guillaud and M. Mirrahimi, Repetition cat qubits for fault-tolerant quantum computation, Phys. Rev. X **9**, 041053 (2019).
[84] A. S. Darmawan, B. J. Brown, A. L. Grimsmo, D. K. Tuckett, and S. Puri, Practical quantum error correction with the XZZX code and Kerr-cat qubits, PRX Quantum **2**, 030345 (2021).
[85] L. B. Nguyen, G. Koolstra, Y. Kim, A. Morvan, T. Chistolini, S. Singh, K. N. Nesterov, C. Jünger, L. Chen, Z. Pedramrazi, B. K. Mitchell, J. M. Kreikebaum, S. Puri, D. I. Santiago, and I. Siddiqi, Blueprint for a high-performance fluxonium quantum processor, PRX Quantum **3**, 037001 (2022).
[86] E. L. Hahn, Spin echoes, Phys. Rev. **80**, 580 (1950).
[87] H. Y. Carr and E. M. Purcell, Effects of diffusion on free precession in nuclear magnetic resonance experiments, Phys. Rev. **94**, 630 (1954).
[88] S. Meiboom and D. Gill, Modified spin-echo method for measuring nuclear relaxation times, Rev. Sci. Instrum. **29**, 688 (1958).
[89] A. Maudsley, Modified Carr-Purcell-Meiboom-Gill sequence for NMR Fourier imaging applications, J. Magn. Reson. (1969) **69**, 488 (1986).
[90] M. A. A. Ahmed, G. A. Alvarez, and D. Suter, Robustness of dynamical decoupling sequences, Phys. Rev. A **87**, 042309 (2013).
[91] C. P. Slichter, *Principles of Magnetic Resonance*, Springer Series in Solid-State Sciences (Springer, Berlin, Germany, 2010)..
[92] A. J. Kerman, Metastable superconducting qubit, Phys. Rev. Lett. **104**, 027002 (2010).
[93] Y.-H. Lin, L. B. Nguyen, N. Grabon, J. San Miguel, N. Pankratova, and V. E. Manucharyan, Demonstration of protection of a superconducting qubit from energy decay, Phys. Rev. Lett. **120**, 150503 (2018).
[94] N. Earnest, S. Chakram, Y. Lu, N. Irons, R. K. Naik, N. Leung, L. Ocola, D. A. Czaplewski, B. Baker, J. Lawrence, J. Koch, and D. I. Schuster, Realization of a Λ system with metastable states of a capacitively shunted fluxonium, Phys. Rev. Lett. **120**, 150504 (2018).
[95] L. B. Nguyen, Y.-H. Lin, A. Somoroff, R. Mencia, N. Grabon, and V. E. Manucharyan, High-coherence fluxonium qubit, Phys. Rev. X **9**, 041041 (2019).
[96] T. Proctor, M. Revelle, E. Nielsen, K. Rudinger, D. Lobser, P. Maunz, R. Blume-Kohout, and K. Young, Detecting and tracking drift in quantum information processors, Nat. Commun. **11**, 1 (2020).
[97] J. Ghosh, A. G. Fowler, J. M. Martinis, and M. R. Geller, Understanding the effects of leakage in superconducting quantum-error-detection circuits, Phys. Rev. A **88**, 062329 (2013).
[98] J. J. Wallman, M. Barnhill, and J. Emerson, Robust characterization of leakage errors, New J. Phys. **18**, 043021 (2016).
[99] Z. Chen, J. Kelly, C. Quintana, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. Fowler, E. Lucero, *et al.*, Measuring and suppressing quantum state leakage

in a superconducting qubit, Phys. Rev. Lett. **116**, 020501 (2016).
[100] C. J. Wood and J. M. Gambetta, Quantification and characterization of leakage errors, Phys. Rev. A **97**, 032306 (2018).
[101] D. Hayes, D. Stack, B. Bjork, A. Potter, C. Baldwin, and R. Stutz, Eliminating leakage errors in hyperfine qubits, Phys. Rev. Lett. **124**, 170501 (2020).
[102] A. P. Babu, J. Tuorila, and T. Ala-Nissila, State leakage during fast decay and control of a superconducting transmon qubit, npj Quantum Inf. **7**, 1 (2021).
[103] X.-G. Li, J.-H. Wang, Y.-Y. Jiang, G.-M. Xue, X.-X. Cai, J. Zhou, M. Gong, Z.-F. Liu, S.-Y. Zheng, D.-K. Ma, *et al.*, Direct evidence for cosmic-ray-induced correlated errors in superconducting qubit array, arXiv:2402.04245.
[104] P. M. Harrington, M. Li, M. Hays, W. Van De Pontseele, D. Mayer, H. D. Pinckney, F. Contipelli, M. Gingras, B. M. Niedzielski, H. Stickler, *et al.*, Synchronous detection of cosmic rays and correlated errors in superconducting qubit arrays, arXiv:2402.03208.
[105] P. Mundada, G. Zhang, T. Hazard, and A. Houck, Suppression of qubit crosstalk in a tunable coupling superconducting circuit, Phys. Rev. Appl. **12**, 054023 (2019).
[106] P. Zhao, P. Xu, D. Lan, J. Chu, X. Tan, H. Yu, and Y. Yu, High-contrast *zz* interaction using superconducting qubits with opposite-sign anharmonicity, Phys. Rev. Lett. **125**, 200503 (2020).
[107] Z. Ni, S. Li, L. Zhang, J. Chu, J. Niu, T. Yan, X. Deng, L. Hu, J. Li, Y. Zhong, *et al.*, Scalable method for eliminating residual *ZZ* interaction between superconducting qubits, Phys. Rev. Lett. **129**, 040502 (2022).
[108] K. Serniak, M. Hays, G. De Lange, S. Diamond, S. Shankar, L. Burkhart, L. Frunzio, M. Houzet, and M. Devoret, Hot nonequilibrium quasiparticles in transmon qubits, Phys. Rev. Lett. **121**, 157701 (2018).
[109] S. de Graaf, L. Faoro, L. Ioffe, S. Mahashabde, J. Burnett, T. Lindström, S. Kubatkin, A. Danilov, and A. Y. Tzalenchuk, Two-level systems in superconducting quantum devices due to trapped quasiparticles, Sci. Adv. **6**, eabc5055 (2020).
[110] M. Berlin-Udi, C. Matthiesen, P. Lloyd, A. Alonso, C. Noel, C. Orme, C.-E. Kim, V. Lordi, and H. Häffner, Changes in electric-field noise due to thermal transformation of a surface ion trap, Phys. Rev. B **106**, 035409 (2022).
[111] A. E. Webb, S. C. Webster, S. Collingbourne, D. Bretaud, A. M. Lawrence, S. Weidt, F. Mintert, and W. K. Hensinger, Resilient entangling gates for trapped ions, Phys. Rev. Lett. **121**, 180501 (2018).
[112] G. Burkard, Non-Markovian qubit dynamics in the presence of $1/f$ noise, Phys. Rev. B **79**, 125317 (2009).
[113] P. Groszkowski, A. Seif, J. Koch, and A. A. Clerk, Simple master equations for describing driven systems subject to classical non-Markovian noise, Quantum **7**, 972 (2023).
[114] L. Diósi, N. Gisin, and W. T. Strunz, Non-Markovian quantum state diffusion, Phys. Rev. A **58**, 1699 (1998).
[115] M. M. Wolf, J. Eisert, T. S. Cubitt, and J. I. Cirac, Assessing non-Markovian quantum dynamics, Phys. Rev. Lett. **101**, 150402 (2008).
[116] J. Piilo, S. Maniscalco, K. Härkönen, and K.-A. Suominen, Non-Markovian quantum jumps, Phys. Rev. Lett. **100**, 180402 (2008).
[117] H.-P. Breuer, E.-M. Laine, and J. Piilo, Measure for the degree of non-Markovian behavior of quantum processes in open systems, Phys. Rev. Lett. **103**, 210401 (2009).
[118] B.-H. Liu, L. Li, Y.-F. Huang, C.-F. Li, G.-C. Guo, E.-M. Laine, H.-P. Breuer, and J. Piilo, Experimental control of the transition from Markovian to non-Markovian dynamics of open quantum systems, Nat. Phys. **7**, 931 (2011).
[119] I. De Vega and D. Alonso, Dynamics of non-Markovian open quantum systems, Rev. Mod. Phys. **89**, 015001 (2017).
[120] J. R. Glick and C. Adami, Markovian and non-Markovian quantum measurements, Found. Phys. **50**, 1008 (2020).
[121] K. Head-Marsden, S. Krastanov, D. A. Mazziotti, and P. Narang, Capturing non-Markovian dynamics on near-term quantum computers, Phys. Rev. Res. **3**, 013182 (2021).
[122] V. Link, W. T. Strunz, and K. Luoma, Non-Markovian quantum dynamics in a squeezed reservoir, Entropy **24**, 352 (2022).
[123] S. Tserkis, K. Head-Marsden, and P. Narang, Information back-flow in quantum non-Markovian dynamics and its connection to teleportation, arXiv:2203.00668.
[124] Á. Rivas, S. F. Huelga, and M. B. Plenio, Quantum non-Markovianity: Characterization, quantification and detection, Rep. Prog. Phys. **77**, 094001 (2014).
[125] H.-P. Breuer, E.-M. Laine, J. Piilo, and B. Vacchini, Colloquium: Non-Markovian dynamics in open quantum systems, Rev. Mod. Phys. **88**, 021002 (2016).
[126] L. Li, M. J. Hall, and H. M. Wiseman, Concepts of quantum non-Markovianity: A hierarchy, Phys. Rep. **759**, 1 (2018).
[127] C.-F. Li, G.-C. Guo, and J. Piilo, Non-Markovian quantum dynamics: What does it mean? EPL (Europhys. Lett.) **127**, 50001 (2019).
[128] S. Milz and K. Modi, Quantum stochastic processes and quantum non-Markovian phenomena, PRX Quantum **2**, 030201 (2021).
[129] G. A. White, P. Jurcevic, C. D. Hill, and K. Modi, Unifying non-Markovian characterisation with an efficient and self-consistent framework, arXiv:2312.08454.
[130] Here, we use the term "metric" loosely. For example, we discuss different types of fidelity in this section, but fidelity is strictly not a metric in a mathematical sense, as it does not obey the triangle inequality.
[131] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distribution, Bull. Calcutta Math. Soc. (1943).
[132] C. A. Fuchs, Distinguishability and accessible information in quantum theory, arXiv:quant-ph/9601020.
[133] C. A. Fuchs and J. Van De Graaf, Cryptographic distinguishability measures for quantum-mechanical states, IEEE Trans. Inf. Theory **45**, 1216 (1999).
[134] The logarithm that appears in entropic quantities can be evaluated in any base; using $\log_2$ yields *bits* of entropy, while ln yields units called *nats*.
[135] Heavy output probability is not necessarily well-defined for highly degenerate distributions.

[136] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Characterizing quantum supremacy in near-term devices, Nat. Phys. **14**, 595 (2018).

[137] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, Validating quantum computers using randomized model circuits, Phys. Rev. A **100**, 032328 (2019).

[138] C. W. Helstrom, Quantum detection and estimation theory, J. Stat. Phys. **1**, 231 (1969).

[139] See, for example, the book *Quantum Computation and Quantum Information* [41].

[140] B. Schumacher, Quantum coding, Phys. Rev. A **51**, 2738 (1995).

[141] A. Uhlmann, The "transition probability" in the state space of a *-algebra, Rep. Math. Phys. **9**, 273 (1976).

[142] R. Jozsa, Fidelity for mixed quantum states, J. Mod. Opt. **41**, 2315 (1994).

[143] It is worth emphasizing that a *quantum process* is not analogous to a classical *stochastic process*. The classical analog of a quantum process is a *stochastic matrix* (see Sec. IX A), which is related to stochastic processes, but quite distinct.

[144] A. Y. Kitaev, Quantum computations: Algorithms and error correction, Uspekhi Matematicheskikh Nauk **52**, 53 (1997).

[145] C. H. Bennett and S. J. Wiesner, Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states, Phys. Rev. Lett. **69**, 2881 (1992).

[146] W. H. Zurek, Environment-assisted invariance, entanglement, and probabilities in quantum physics, Phys. Rev. Lett. **90**, 120404 (2003).

[147] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels, Phys. Rev. Lett. **70**, 1895 (1993).

[148] D. Aharonov, A. Kitaev, and N. Nisan, in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 1998), p. 20.

[149] L. Viola and S. Lloyd, Dynamical suppression of decoherence in two-state quantum systems, Phys. Rev. A **58**, 2733 (1998).

[150] L. Viola, E. Knill, and S. Lloyd, Dynamical decoupling of open quantum systems, Phys. Rev. Lett. **82**, 2417 (1999).

[151] E. Knill, Fault-tolerant postselected quantum computation: Threshold analysis, arXiv:quant-ph/0404104.

[152] O. Kern, G. Alber, and D. L. Shepelyansky, Quantum error correction of coherent errors by randomization, Eur. Phys. J. D-At. Mol. Opt. Plasma Phys. **32**, 153 (2005).

[153] M. Ware, G. Ribeill, D. Riste, C. A. Ryan, B. Johnson, and M. P. Da Silva, Experimental Pauli-frame randomization on a superconducting qubit, Phys. Rev. A **103**, 042604 (2021).

[154] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, Phys. Rev. A **94**, 052325 (2016).

[155] A. Hashim, R. K. Naik, A. Morvan, J.-L. Ville, B. Mitchell, J. M. Kreikebaum, M. Davis, E. Smith, C. Iancu, K. P. O'Brien, I. Hincks, J. J. Wallman, J. Emerson, and I. Siddiqi, Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor, Phys. Rev. X **11**, 041039 (2021).

[156] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout, Gate set tomography, Quantum **5**, 557 (2021).

[157] P. Aliferis, D. Gottesman, and J. Preskill, Quantum accuracy threshold for concatenated distance-3 codes, Quantum Inf. Comput. **6**, 97 (2006).

[158] Y. R. Sanders, J. J. Wallman, and B. C. Sanders, Bounding quantum gate error rate based on reported average fidelity, New J. Phys. **18**, 012002 (2015).

[159] M. A. Nielsen, A simple formula for the average gate fidelity of a quantum dynamical operation, Phys. Lett. A **303**, 249 (2002).

[160] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B: Quantum Semiclassical Opt. **7**, S347 (2005).

[161] E. Magesan, R. Blume-Kohout, and J. Emerson, Gate fidelity fluctuations and quantum process invariants, Phys. Rev. A **84**, 012309 (2011).

[162] M. Horodecki, P. Horodecki, and R. Horodecki, General teleportation channel, singlet fraction, and quasidistillation, Phys. Rev. A **60**, 1888 (1999).

[163] B. Schumacher, Sending entanglement through noisy quantum channels, Phys. Rev. A **54**, 2614 (1996).

[164] M. A. Nielsen, The entanglement fidelity and quantum error correction, arXiv:quant-ph/9606012.

[165] A. Gilchrist, N. K. Langford, and M. A. Nielsen, Distance measures to compare real and ideal quantum processes, Phys. Rev. A **71**, 062310 (2005).

[166] A. Carignan-Dugas, A walk through quantum noise: A study of error signatures and characterization methods, Ph.D. thesis, University of Waterloo, 2019.

[167] R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz, Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography, Nat. Commun. **8**, 14485 (2017).

[168] J. Watrous, Semidefinite programs for completely bounded norms, Theory Comput. **5**, 217 (2009).

[169] J. Watrous, Simpler semidefinite programs for completely bounded norms, arXiv:1207.5726.

[170] J. J. Wallman and S. T. Flammia, Randomized benchmarking with confidence, New J. Phys. **16**, 103032 (2014).

[171] J. J. Wallman, Bounding experimental quantum error rates relative to fault-tolerant thresholds, arXiv:1511.00727.

[172] R. Kueng, D. M. Long, A. C. Doherty, and S. T. Flammia, Comparing experiments to the fault-tolerance threshold, Phys. Rev. Lett. **117**, 170502 (2016).

[173] A. Luis and L. L. Sánchez-Soto, Complete characterization of arbitrary quantum measurement processes, Phys. Rev. Lett. **83**, 3573 (1999).

[174] Z. Ji, Y. Feng, R. Duan, and M. Ying, Identification and distance measures of measurement apparatus, Phys. Rev. Lett. **96**, 200401 (2006).

[175] E. Magesan and P. Cappellaro, Experimentally efficient methods for estimating the performance of quantum measurements, Phys. Rev. A **88**, 022127 (2013).

[176] J. Dressel, T. A. Brun, and A. N. Korotkov, Implementing generalized measurements with superconducting qubits, Phys. Rev. A **90**, 032302 (2014).

[177] J. Z. Blumoff, K. Chou, C. Shen, M. Reagor, C. Axline, R. Brierley, M. Silveri, C. Wang, B. Vlastakis, S. E. Nigg, *et al.*, Implementing and characterizing precise multiqubit measurements, Phys. Rev. X **6**, 031041 (2016).

[178] F. Mallet, F. R. Ong, A. Palacios-Laloy, F. Nguyen, P. Bertet, D. Vion, and D. Esteve, Single-shot qubit readout in circuit quantum electrodynamics, Nat. Phys. **5**, 791 (2009).

[179] J. Johnson, C. Macklin, D. Slichter, R. Vijay, E. Weingarten, J. Clarke, and I. Siddiqi, Heralded state preparation in a superconducting qubit, Phys. Rev. Lett. **109**, 050506 (2012).

[180] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, *et al.*, Rapid high-fidelity multiplexed readout of superconducting qubits, Phys. Rev. Appl. **10**, 034040 (2018).

[181] S. S. Elder, C. S. Wang, P. Reinhold, C. T. Hann, K. S. Chou, B. J. Lester, S. Rosenblum, L. Frunzio, L. Jiang, and R. J. Schoelkopf, High-fidelity measurement of qubits encoded in multilevel superconducting circuits, Phys. Rev. X **10**, 011001 (2020).

[182] S. J. Beale and J. J. Wallman, Randomized compiling for subsystem measurements, arXiv:2304.06599.

[183] A. Hashim, A. Carignan-Dugas, L. Chen, C. Juenger, N. Fruitwala, Y. Xu, G. Huang, J. Wallman, and I. Siddiqi, Quasi-probabilistic readout correction of mid-circuit measurements for adaptive feedback via measurement randomized compiling, PRX Quantum **6**, 010307 (2025).

[184] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. Da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, *et al.*, Efficient measurement of quantum gate error by interleaved randomized benchmarking, Phys. Rev. Lett. **109**, 080505 (2012).

[185] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Symmetric informationally complete quantum measurements, J. Math. Phys. **45**, 2171 (2004).

[186] J. M. Chow, J. M. Gambetta, E. Magesan, D. W. Abraham, A. W. Cross, B. R. Johnson, N. A. Masluk, C. A. Ryan, J. A. Smolin, S. J. Srinivasan, and M. Steffen, Implementing a strand of a scalable fault-tolerant quantum computing fabric, Nat. Commun. **5**, 4015 (2014).

[187] A. S. Holevo, Quantum coding theorems, Russ. Math. Surv. **53**, 1295 (1998).

[188] D. McLaren, M. A. Graydon, and J. J. Wallman, Stochastic errors in quantum instruments, arXiv:2306.07418.

[189] L. Pereira, J. J. García-Ripoll, and T. Ramos, Complete physical characterization of quantum nondemolition measurements via tomography, Phys. Rev. Lett. **129**, 010402 (2022).

[190] L. Pereira, J. J. García-Ripoll, and T. Ramos, Parallel tomography of quantum non-demolition measurements in multi-qubit devices, npj Quantum Inf. **9**, 22 (2023).

[191] Y. Nakamura, Y. A. Pashkin, and J. S. Tsai, Coherent control of macroscopic quantum states in a single-Cooper-pair box, Nature **398**, 786 (1999).

[192] O. Astafiev, A. M. Zagoskin, A. A. Abdumalikov, Y. A. Pashkin, T. Yamamoto, K. Inomata, Y. Nakamura, and J. S. Tsai, Resonance fluorescence of a single artificial atom, Science **327**, 840 (2010).

[193] N. Cottet, H. Xiong, L. B. Nguyen, Y.-H. Lin, and V. E. Manucharyan, Electron shelving of a superconducting artificial atom, Nat. Commun. **12**, 1 (2021).

[194] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Self-consistent quantum process tomography, Phys. Rev. A **87**, 062119 (2013).

[195] R. Blume-Kohout, J. K. Gamble, E. Nielsen, J. Mizrahi, J. D. Sterk, and P. Maunz, Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit, arXiv:1310.4492.

[196] T. Proctor, K. Rudinger, K. Young, M. Sarovar, and R. Blume-Kohout, What randomized benchmarking actually measures, Phys. Rev. Lett. **119**, 130502 (2017).

[197] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout, Gate set tomography, Quantum **5**, 557 (2021).

[198] J. J. Wallman, Randomized benchmarking with gate-dependent noise, Quantum **2**, 47 (2018).

[199] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

[200] S. Dasgupta and T. S. Humble, Characterizing the reproducibility of noisy quantum circuits, Entropy **24**, 244 (2022).

[201] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Quantum state tomography via compressed sensing, Phys. Rev. Lett. **105**, 150401 (2010).

[202] C. A. Riofrío, D. Gross, S. T. Flammia, T. Monz, D. Nigg, R. Blatt, and J. Eisert, Experimental quantum compressed sensing for a seven-qubit system, Nat. Commun. **8**, 15305 (2017).

[203] J. J. Meyer, Fisher information in noisy intermediate-scale quantum applications, Quantum **5**, 539 (2021).

[204] C. Ostrove, K. Rudinger, S. Seritan, K. Young, and R. Blume-Kohout, in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, Vol. 1 (IEEE, 2023), p. 1422.

[205] C. Tong, H. Zhang, and B. Pokharel, Empirical learning of dynamical decoupling on quantum processors, arXiv:2403.02294.

[206] T. Proctor, S. Seritan, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, Scalable randomized benchmarking of quantum computers using mirror circuits, Phys. Rev. Lett. **129**, 150502 (2022).

[207] G. Torlai, C. J. Wood, A. Acharya, G. Carleo, J. Carrasquilla, and L. Aolita, Quantum process tomography with unsupervised learning and tensor networks, Nat. Commun. **14**, 2858 (2023).

[208] T. J. Evans, R. Harper, and S. T. Flammia, Scalable Bayesian Hamiltonian learning, arXiv:1912.07636.

[209] H.-Y. Huang, R. Kueng, and J. Preskill, Efficient estimation of Pauli observables by derandomization, Phys. Rev. Lett. **127**, 030503 (2021).

[210] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[211] R. W. Andrews, C. Jones, M. D. Reed, A. M. Jones, S. D. Ha, M. P. Jura, J. Kerckhoff, M. Levendorf, S. Meenehan, S. T. Merkel, *et al.*, Quantifying error and leakage in an

encoded Si/SiGe triple-dot qubit, Nat. Nanotechnol. **14**, 747 (2019).
[212] J. Helsen, X. Xue, L. M. Vandersypen, and S. Wehner, A new class of efficient randomized benchmarking protocols, npj Quantum Inf. **5**, 71 (2019).
[213] R. S. Gupta, L. C. G. Govia, and M. J. Biercuk, Integration of spectator qubits into quantum computer architectures for hardware tune-up and calibration, Phys. Rev. A **102**, 042611 (2020).
[214] S. Majumder, L. A. de Castro, and K. R. Brown, Real-time calibration with spectator qubits, npj Quantum Inf. **6**, 1 (2020).
[215] N. Fruitwala, A. Hashim, A. D. Rajagopala, Y. Xu, J. Hines, R. K. Naik, I. Siddiqi, K. Klymko, G. Huang, and K. Nowrouzi, Hardware-efficient randomized compiling, arXiv:2406.13967.
[216] M. J. Schervish, P values: What they are and what they are not, Am. Stat. **50**, 203 (1996).
[217] R. Blume-Kohout, K. Rudinger, E. Nielsen, T. Proctor, and K. Young, Wildcard error: Quantifying unmodeled errors in quantum processors, arXiv:2012.12231.
[218] L. B. Nguyen, Y. Kim, A. Hashim, N. Goss, B. Marinelli, B. Bhandari, D. Das, R. K. Naik, J. M. Kreikebaum, A. N. Jordan, *et al.*, Programmable Heisenberg interactions between Floquet qubits, Nat. Phys. **20**, 240 (2024).
[219] W. Demtröder, *Laser Spectroscopy*, Vol. 2 (Springer, 1973), https://link.springer.com/book/10.1007/3-540-06334-X.
[220] N. F. Ramsey, A molecular beam resonance method with separated oscillating fields, Phys. Rev. **78**, 695 (1950).
[221] A. Somoroff, Q. Ficheux, R. A. Mencia, H. Xiong, R. Kuzmin, and V. E. Manucharyan, Millisecond coherence in a superconducting qubit, Phys. Rev. Lett. **130**, 267001 (2023).
[222] P. Wang, C.-Y. Luan, M. Qiao, M. Um, J. Zhang, Y. Wang, X. Yuan, M. Gu, J. Zhang, and K. Kim, Single ion qubit with estimated coherence time exceeding one hour, Nat. Commun. **12**, 233 (2021).
[223] A. G. Redfield, On the theory of relaxation processes, IBM J. Res. Dev. **1**, 19 (1957).
[224] The term "phase estimation" has multiple and closely related uses, as it also refers to the similar but distinct task of estimating a Hamiltonian's eigenvalues [19].
[225] S. Kimmel, G. H. Low, and T. J. Yoder, Robust calibration of a universal single-qubit gate set via robust phase estimation, Phys. Rev. A **92**, 062315 (2015).
[226] A unit-normalized linear combination of Pauli matrices also works, i.e., $\hat{\mathbf{n}} \cdot \boldsymbol{\sigma}$, but makes state preparation and measurement mildly more complicated.
[227] A. E. Russo, W. M. Kirby, K. M. Rudinger, A. D. Baczewski, and S. Kimmel, Consistency testing for robust phase estimation, Phys. Rev. A **103**, 042609 (2021).
[228] K. Rudinger, S. Kimmel, D. Lobser, and P. Maunz, Experimental demonstration of a cheap and accurate phase estimation, Phys. Rev. Lett. **118**, 190502 (2017).
[229] pyrpe, https://https://gitlab.com/quapack/pyrpe (2021).
[230] A. E. Russo, K. M. Rudinger, B. C. A. Morrison, and A. D. Baczewski, Evaluating energy differences on a quantum computer with robust phase estimation, Phys. Rev. Lett. **126**, 210501 (2021).
[231] K. Rudinger, J. P. Marceaux, A. Hashim, D. I. Santiago, I. Siddiqi, and K. C. Young, Heisenberg-limited calibration of entangling gates with robust phase estimation, arXiv:2502.06698.
[232] W. Gale, E. Guth, and G. Trammell, Determination of the quantum state by measurements, Phys. Rev. **165**, 1434 (1968).
[233] I. L. Chuang and M. A. Nielsen, Prescription for experimental determination of the dynamics of a quantum black box, J. Mod. Opt. **44**, 2455 (1997).
[234] W. K. Wootters and B. D. Fields, Optimal state-determination by mutually unbiased measurements, Ann. Phys. **191**, 363 (1989).
[235] R. Adamson and A. M. Steinberg, Improving quantum state estimation with mutually unbiased bases, Phys. Rev. Lett. **105**, 030406 (2010).
[236] R. Stricker, M. Meth, L. Postler, C. Edmunds, C. Ferrie, R. Blatt, P. Schindler, T. Monz, R. Kueng, and M. Ringbauer, Experimental single-setting quantum state tomography, PRX Quantum **3**, 040310 (2022).
[237] S. T. Flammia, A. Silberfarb, and C. M. Caves, Minimal informationally complete measurements for pure states, Found. Phys. **35**, 1985 (2005).
[238] S. Filippov and V. Man'Ko, Mutually unbiased bases: Tomography of spin states and the star-product scheme, Phys. Scr. **2011**, 014010 (2011).
[239] D. E. Gottesman, Stabilizer codes and quantum error correction, Ph.D. thesis, California Institute of Technology, 1997.
[240] Shot noise is not the only source of fluctuations and errors. Laboratory measurements are also subject to, for example, imperfect signal amplification, electronic noise, poor quantum efficiency, imperfect signal conversion, errors in digitization and classification, and a terrifying range of *systematic* errors like drift over the duration of a tomography experiment. However, there is no systematic theoretical treatment of these noise sources. In practice, the techniques used to deal with shot noise (which *does* have a solid theory) can deal with these noise sources too, although not optimally.
[241] J. A. Smolin, J. M. Gambetta, and G. Smith, Efficient method for computing the maximum likelihood quantum state from measurements with additive Gaussian noise, Phys. Rev. Lett. **108**, 070502 (2012).
[242] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp, Fast state tomography with optimal error bounds, J. Phys. A: Math. Theor. **53**, 204001 (2020).
[243] T. Surawy-Stepney, J. Kahn, R. Kueng, and M. Guta, Projected least-squares quantum process tomography, Quantum **6**, 844 (2022).
[244] Z. Hradil, Quantum-state estimation, Phys. Rev. A **55**, R1561 (1997).
[245] K. Banaszek, G. M. D'Ariano, M. G. A. Paris, and M. F. Sacchi, Maximum-likelihood estimation of the density matrix, Phys. Rev. A **61**, 010304 (1999).
[246] R. Blume-Kohout, Optimal, reliable estimation of quantum states, New J. Phys. **12**, 043034 (2010).
[247] V. Bužek, R. Derka, G. Adam, and P. L. Knight, Reconstruction of quantum states of spin systems: From quantum Bayesian inference to quantum tomography, Ann. Phys. **266**, 454 (1998).

[248] C. Granade, J. Combes, and D. Cory, Practical Bayesian tomography, New J. Phys. **18**, 033024 (2016).
[249] C. M. Caves, C. A. Fuchs, and R. Schack, Quantum probabilities as Bayesian probabilities, Phys. Rev. A **65**, 022305 (2002).
[250] U. Von Toussaint, Bayesian inference in physics, Rev. Mod. Phys. **83**, 943 (2011).
[251] S. M. Barnett and S. Croke, Quantum state discrimination, Adv. Opt. Photonics **1**, 238 (2009).
[252] J. Bae and L.-C. Kwek, Quantum state discrimination and its applications, J. Phys. A: Math. Theor. **48**, 083001 (2015).
[253] V. Siddhu, Maximum a posteriori probability estimates for quantum tomography, Phys. Rev. A **99**, 012342 (2019).
[254] R. Blume-Kohout, Hedged maximum likelihood quantum state estimation, Phys. Rev. Lett. **105**, 200504 (2010).
[255] F. Huszár and N. M. Houlsby, Adaptive Bayesian quantum tomography, Phys. Rev. A–At. Mol. Opt. Phys. **85**, 052120 (2012).
[256] K. S. Kravtsov, S. S. Straupe, I. V. Radchenko, N. M. Houlsby, F. Huszár, and S. P. Kulik, Experimental adaptive Bayesian tomography, Phys. Rev. A–At. Mol. Opt. Phys. **87**, 062122 (2013).
[257] C. Granade, C. Ferrie, and S. T. Flammia, Practical adaptive quantum tomography, New J. Phys. **19**, 113017 (2017).
[258] M. Christandl and R. Renner, Reliable quantum state tomography, Phys. Rev. Lett. **109**, 120403 (2012).
[259] C. Oh, Y. S. Teo, and H. Jeong, Efficient Bayesian credible-region certification for quantum-state tomography, Phys. Rev. A **100**, 012345 (2019).
[260] J. M. Lukens, K. J. Law, A. Jasra, and P. Lougovski, A practical and efficient approach for Bayesian quantum state estimation, New J. Phys. **22**, 063038 (2020).
[261] T. Evans, W. Huang, J. Yoneda, R. Harper, T. Tanttu, K. Chan, F. Hudson, K. Itoh, A. Saraiva, C. Yang, *et al.*, Fast Bayesian tomography of a two-qubit gate set in silicon, Phys. Rev. Appl. **17**, 024068 (2022).
[262] J. Poyatos, J. I. Cirac, and P. Zoller, Complete characterization of a quantum process: The two-bit quantum gate, Phys. Rev. Lett. **78**, 390 (1997).
[263] Y. Kim, A. Morvan, L. B. Nguyen, R. K. Naik, C. Jünger, L. Chen, J. M. Kreikebaum, D. I. Santiago, and I. Siddiqi, High-fidelity three-qubit iToffoli gate for fixed-frequency superconducting qubits, Nat. Phys. **18**, 783 (2022).
[264] J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, M. B. Ketchen, and M. Steffen, Universal quantum gate set approaching fault-tolerant thresholds with superconducting qubits, Phys. Rev. Lett. **109**, 060501 (2012).
[265] A. D. Córcoles, J. M. Gambetta, J. M. Chow, J. A. Smolin, M. Ware, J. Strand, B. L. T. Plourde, and M. Steffen, Process verification of two-qubit quantum gates by randomized benchmarking, Phys. Rev. A **87**, 030301 (2013).
[266] R. Blume-Kohout and T. Proctor, Easy better quantum process tomography, arXiv:2412.16293.
[267] M. Mitchell, C. Ellenor, S. Schneider, and A. Steinberg, Diagnosis, prescription, and prognosis of a bell-state filter by quantum process tomography, Phys. Rev. Lett. **91**, 120402 (2003).
[268] J. L. O'Brien, G. J. Pryde, A. Gilchrist, D. F. James, N. K. Langford, T. C. Ralph, and A. G. White, Quantum process tomography of a controlled-not gate, Phys. Rev. Lett. **93**, 080502 (2004).
[269] G. C. Knee, E. Bolduc, J. Leach, and E. M. Gauger, Quantum process tomography via completely positive and trace-preserving projection, Phys. Rev. A **98**, 062336 (2018).
[270] J. Fiurášek, Maximum-likelihood estimation of quantum measurement, Phys. Rev. A **64**, 024102 (2001).
[271] J. S. Lundeen, A. Feito, H. Coldenstrodt-Ronge, K. L. Pregnell, C. Silberhorn, T. C. Ralph, J. Eisert, M. B. Plenio, and I. A. Walmsley, Tomography of quantum detectors, Nat. Phys. **5**, 27 (2009).
[272] L. B. Nguyen, N. Goss, K. Siva, Y. Kim, E. Younis, B. Qing, A. Hashim, D. I. Santiago, and I. Siddiqi, Empowering high-dimensional quantum computing by traversing the dual bosonic ladder, Nat. Commun. **15**, 7117 (2024).
[273] S. Bravyi, S. Sheldon, A. Kandala, D. C. Mckay, and J. M. Gambetta, Mitigating measurement errors in multiqubit experiments, Phys. Rev. A **103**, 042605 (2021).
[274] Y. Chen, M. Farahzad, S. Yoo, and T.-C. Wei, Detector tomography on IBM quantum computers and mitigation of an imperfect measurement, Phys. Rev. A **100**, 052315 (2019).
[275] D. Greenbaum, Introduction to quantum gate set tomography, arXiv:1509.02921.
[276] Y. Gu, R. Mishra, B.-G. Englert, and H. K. Ng, Randomized linear gate set tomography, PRX Quantum **2**, 030328 (2021).
[277] J. P. Dehollain, J. T. Muhonen, R. Blume-Kohout, K. M. Rudinger, J. K. Gamble, E. Nielsen, A. Laucht, S. Simmons, R. Kalra, A. S. Dzurak, *et al.*, Optimization of a solid-state electron spin qubit using gate set tomography, New J. Phys. **18**, 103018 (2016).
[278] X. Xue, M. Russ, N. Samkharadze, B. Undseth, A. Sammak, G. Scappucci, and L. M. K. Vandersypen, Quantum logic with spin qubits crossing the surface code threshold, Nature **601**, 343 (2022).
[279] E. Nielsen, K. Rudinger, T. Proctor, A. Russo, K. Young, and R. Blume-Kohout, Probing quantum processor performance with pyGSTi, Quantum Sci. Technol. **5**, 044002 (2020).
[280] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum-enhanced measurements: Beating the standard quantum limit, Science **306**, 1330 (2004).
[281] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. **9**, 60 (1938).
[282] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).
[283] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).

[284] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and robust randomized benchmarking of quantum processes, Phys. Rev. Lett. **106**, 180504 (2011).

[285] Here, we define a circuit of depth $m = 0$ to be the minimal benchmark depth, which contains only a single random gate (and its inverse). By defining it this way, the error in any gates in the $m = 0$ circuit all contributes to effective SPAM error. Therefore, any gates used for state preparation or basis rotations for measurement can be compiled into the initial and final circuit layers, respectively.

[286] D. Gottesman, The Heisenberg representation of quantum computers, arXiv:quant-ph/9807006.

[287] R. Koenig and J. A. Smolin, How to efficiently select an arbitrary Clifford group element, J. Math. Phys. **55**, 122202 (2014).

[288] M. A. Fogarty, M. Veldhorst, R. Harper, C. Yang, S. Bartlett, S. T. Flammia, and A. Dzurak, Nonexponential fidelity decay in randomized benchmarking with low-frequency noise, Phys. Rev. A **92**, 022326 (2015).

[289] J. T. Muhonen, A. Laucht, S. Simmons, J. P. Dehollain, R. Kalra, F. E. Hudson, S. Freer, K. M. Itoh, D. N. Jamieson, J. C. McCallum, *et al.*, Quantifying the quantum gate fidelity of single-atom spin qubits in silicon by randomized benchmarking, J. Phys.: Condens. Matter **27**, 154205 (2015).

[290] R. Harper, I. Hincks, C. Ferrie, S. T. Flammia, and J. J. Wallman, Statistical analysis of randomized benchmarking, Phys. Rev. A **99**, 052350 (2019).

[291] This ensures that the entire circuit is not compiled down into a single gate layer, which would defeat the purpose of the benchmark.

[292] C. Granade, C. Ferrie, and D. G. Cory, Accelerated randomized benchmarking, New J. Phys. **17**, 013042 (2015).

[293] $Ap^m + B$ is the success probability of circuits containing $m + 2$ Clifford gates and a SPAM operation, so $Ap^{-2} + B$, i.e., extrapolating the decay curve back to $m = -2$, is a heuristic for SPAM error's contribution to the success probabilities deviation from 1. Note that A is not invariant under changes in the convention for "depth" in RB circuits—e.g., defining m to be the number of uniformly random Clifford gates in the circuit—but this heuristic for SPAM fidelity is.

[294] T. Proctor, K. Rudinger, K. Young, E. Nielsen, and R. Blume-Kohout, Measuring the capabilities of quantum computers, Nat. Phys. **18**, 75 (2022).

[295] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, Nature **508**, 500 (2014).

[296] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, Efficient z gates for quantum computing, Phys. Rev. A **96**, 022330 (2017).

[297] Virtual Z gates do not implement physical pulses; rather, they provide a frame update (i.e., a shift in phase) for the subsequent physical pulse.

[298] A. Morvan, V. Ramasesh, M. Blok, J. Kreikebaum, K. O'Brien, L. Chen, B. Mitchell, R. Naik, D. Santiago, and I. Siddiqi, Qutrit randomized benchmarking, Phys. Rev. Lett. **126**, 210504 (2021).

[299] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A. **85**, 042311 (2012).

[300] S. T. Merkel, E. J. Pritchett, and B. H. Fong, Randomized benchmarking as convolution: Fourier analysis of gate dependent errors, Quantum **5**, 581 (2021).

[301] J. Helsen, I. Roth, E. Onorati, A. Werner, and J. Eisert, General framework for randomized benchmarking, PRX Quantum **3**, 020357 (2022).

[302] A. Carignan-Dugas, K. Boone, J. J. Wallman, and J. Emerson, From randomized benchmarking experiments to gate set circuit fidelity: How to interpret randomized benchmarking decay parameters, New J. Phys. **20**, 092001 (2018).

[303] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).

[304] D. Maslov and M. Roetteler, Shorter stabilizer circuits via Bruhat decomposition and quantum circuit transformations, IEEE Trans. Inf. Theory **64**, 4729 (2018).

[305] S. Bravyi and D. Maslov, Hadamard-free circuits expose the structure of the Clifford group, IEEE Trans. Inf. Theory **67**, 4546 (2021).

[306] T. Proctor and K. Young, A simple asymptotically optimal Clifford circuit compilation algorithm, arXiv:2310.10882.

[307] K. N. Patel, I. L. Markov, and J. P. Hayes, Optimal synthesis of linear reversible circuits, Quantum Inf. Comput. **8**, 282 (2008).

[308] A. M. Polloreno, A. Carignan-Dugas, J. Hines, R. Blume-Kohout, K. Young, and T. Proctor, A theory of direct randomized benchmarking, arXiv:2302.13853.

[309] J. M. Epstein, A. W. Cross, E. Magesan, and J. M. Gambetta, Investigating the limits of randomized benchmarking protocols, Phys. Rev. A **89**, 062321 (2014).

[310] T. J. Proctor, A. Carignan-Dugas, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, Direct randomized benchmarking for multiqubit devices, Phys. Rev. Lett. **123**, 030503 (2019).

[311] J. Hines, M. Lu, R. K. Naik, A. Hashim, J.-L. Ville, B. Mitchell, J. M. Kriekebaum, D. I. Santiago, S. Seritan, E. Nielsen, R. Blume-Kohout, K. Young, I. Siddiqi, B. Whaley, and T. Proctor, Demonstrating scalable randomized benchmarking of universal gate sets, Phys. Rev. X **13**, 041030 (2023).

[312] J. Hines, D. Hothem, R. Blume-Kohout, B. Whaley, and T. Proctor, Fully scalable randomized benchmarking without motion reversal, PRX Quantum **5**, 030334 (2024).

[313] Here, a "success metric" does not necessarily refer to a "metric" in the strict mathematical sense; see the discussion in Sec. IV.

[314] D. C. McKay, I. Hincks, E. J. Pritchett, M. Carroll, L. C. Govia, and S. T. Merkel, Benchmarking quantum processor performance at scale, arXiv:2311.05933.

[315] J.-S. Chen, E. Nielsen, M. Ebert, V. Inlek, K. Wright, V. Chaplin, A. Maksymov, E. Páez, A. Poudel, P. Maunz, *et al.*, Benchmarking a trapped-ion quantum computer with 30 qubits, Quantum **8**, 1516 (2024).

[316] K. Mayer, A. Hall, T. Gatterman, S. K. Halit, K. Lee, J. Bohnet, D. Gresh, A. Hankin, K. Gilmore, J. Gerber, *et al.*,

Theory of mirror benchmarking and demonstration on a quantum computer, arXiv:2108.10431.
[317] M. Amico, H. Zhang, P. Jurcevic, L. S. Bishop, P. Nation, A. Wack, and D. C. McKay, Defining standard strategies for quantum benchmarks, arXiv:2303.02108.
[318] T. Proctor, S. Seritan, E. Nielsen, K. Rudinger, K. Young, R. Blume-Kohout, and M. Sarovar, Establishing trust in quantum computations, arXiv:2204.07568.
[319] C. Neill, *et al.*, A blueprint for demonstrating quantum supremacy with superconducting qubits, Science **360**, 195 (2018).
[320] Y. Liu, M. Otten, R. Bassirianjahromi, L. Jiang, and B. Fefferman, Benchmarking near-term quantum computers via random circuit sampling, arXiv:2105.05232.
[321] M. Heinrich, M. Kliesch, and I. Roth, Randomized benchmarking with random quantum circuits, arXiv:2212.06181.
[322] J. Chen, D. Ding, C. Huang, and L. Kong, Linear cross-entropy benchmarking with Clifford circuits, Phys. Rev. A **108**, 052613 (2023).
[323] B. Ware, A. Deshpande, D. Hangleiter, P. Niroula, B. Fefferman, A. V. Gorshkov, and M. J. Gullans, A sharp phase transition in linear cross-entropy benchmarking, arXiv:2305.04954.
[324] Because XEB requires that an n-qubit circuit converges to an n-qubit Haar random unitary, the infidelity of twirling layers consisting only of Haar random single-qubit gates cannot be measured via an n-qubit XEB experiment; rather, it must be estimated from the combined infidelity of simultaneous XEB on all n qubits. Or, instead, one could use n-qubit Haar random unitaries for the twirl, in which case an n-qubit XEB experiment without the interleaved gate could be used to estimate the infidelity of the twirling layer. However, in this case, the decomposition of XEB circuits to native gates would scale poorly (similar to n-qubit CRB). Furthermore, note that the estimate of the interleaved gate's fidelity would be subject to systematic errors similar to those seen in IRB.
[325] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, Real randomized benchmarking, Quantum **2**, 85 (2018).
[326] W. G. Brown and B. Eastin, Randomized benchmarking with restricted gate sets, Phys. Rev. A **97**, 062323 (2018).
[327] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, Characterizing universal gate sets via dihedral benchmarking, Phys. Rev. A **92**, 060302 (2015).
[328] J. Claes, E. Rieffel, and Z. Wang, Character randomized benchmarking for non-multiplicity-free groups with applications to subspace, leakage, and matchgate randomized benchmarking, PRX Quantum **2**, 010351 (2021).
[329] J. Helsen, S. Nezami, M. Reagor, and M. Walter, Matchgate benchmarking: Scalable benchmarking of a continuous family of many-qubit gates, Quantum **6**, 657 (2022).
[330] D. S. França and A. K. Hashagen, Approximate randomized benchmarking for finite groups, J. Phys. A: Math. Theor. **51**, 395302 (2018).
[331] J. Claes and S. Puri, Estimating the bias of CX gates via character randomized benchmarking, PRX Quantum **4**, 010307 (2023).
[332] J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, M. B. Ketchen, and M. Steffen, Characterization of addressability by simultaneous randomized benchmarking, Phys. Rev. Lett. **109**, 240504 (2012).
[333] D. C. McKay, S. Sheldon, J. A. Smolin, J. M. Chow, and J. M. Gambetta, Three-qubit randomized benchmarking, Phys. Rev. Lett. **122**, 200502 (2019).
[334] D. C. McKay, A. W. Cross, C. J. Wood, and J. M. Gambetta, Correlated randomized benchmarking, arXiv:2003.02354.
[335] R. Harper, S. T. Flammia, and J. J. Wallman, Efficient learning of quantum noise, Nat. Phys. **16**, 1184 (2020).
[336] R. Harper and S. T. Flammia, Learning correlated noise in a 39-qubit quantum processor, PRX Quantum **4**, 040311 (2023).
[337] S. Garion, N. Kanazawa, H. Landa, D. C. McKay, S. Sheldon, A. W. Cross, and C. J. Wood, Experimental implementation of non-Clifford interleaved randomized benchmarking with a controlled-s gate, Phys. Rev. Res. **3**, 013204 (2021).
[338] R. Harper and S. T. Flammia, Estimating the fidelity of T gates using standard interleaved randomized benchmarking, Quantum Sci. Technol. **2**, 015008 (2017).
[339] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, Bounding the average gate fidelity of composite channels using the unitarity, New J. Phys. **21**, 053016 (2019).
[340] D. Hothem, J. Hines, K. Nataraj, R. Blume-Kohout, and T. Proctor, in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, Vol. 1 (IEEE Computer Society, Los Alamitos, CA, USA, 2023), p. 709.
[341] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, Characterizing large-scale quantum computers via cycle benchmarking, Nat. Commun. **10**, 1 (2019).
[342] S. J. Beale, A. Carignan-Dugas, D. Dahlen, J. Emerson, I. Hincks, P. Iyer, A. Jain, D. Hufnagel, E. Ospadov, J. Saunders, A. Stasiuk, J. J. Wallman, and A. Winick, Trueq, 2020.
[343] B. K. Mitchell, R. K. Naik, A. Morvan, A. Hashim, J. M. Kreikebaum, B. Marinelli, W. Lavrijsen, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, Hardware-efficient microwave-activated tunable coupling between superconducting qubits, Phys. Rev. Lett. **127**, 200502 (2021).
[344] A. Hashim, R. Rines, V. Omole, R. K. Naik, J. M. Kreikebaum, D. I. Santiago, F. T. Chong, I. Siddiqi, and P. Gokhale, Optimized swap networks with equivalent circuit averaging for QAOA, Phys. Rev. Res. **4**, 033028 (2022).
[345] By extension, it can also be used to measure the process fidelity of an entire subcircuit, and can therefore be considered a form of SPAM-robust fidelity estimation; see Sec. X.
[346] S. Krinner, S. Lazar, A. Remm, C. Andersen, N. Lacroix, G. Norris, C. Hellings, M. Gabureac, C. Eichler, and A. Wallraff, Benchmarking coherent errors in controlled-phase gates due to spectator qubits, Phys. Rev. Appl. **14**, 024042 (2020).
[347] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, Estimating the coherence of noise, New J. Phys. **17**, 113020 (2015).

[348] G. Feng, J. J. Wallman, B. Buonacorsi, F. H. Cho, D. K. Park, T. Xin, D. Lu, J. Baugh, and R. Laflamme, Estimating the coherence of noise in quantum control of a solid-state qubit, Phys. Rev. Lett. **117**, 260501 (2016).

[349] A. Zhu, J. H. Béjanin, X. Xu, and M. Mariantoni, Purity benchmarking study of error coherence in a single xmon qubit, arXiv:2407.07960.

[350] S. Sheldon, L. S. Bishop, E. Magesan, S. Filipp, J. M. Chow, and J. M. Gambetta, Characterizing errors on qubit operations via iterative randomized benchmarking, Phys. Rev. A **93**, 012301 (2016).

[351] I. N. Moskalenko, I. A. Simakov, N. N. Abramov, A. A. Grigorev, D. O. Moskalev, A. A. Pishchimova, N. S. Smirnov, E. V. Zikiy, I. A. Rodionov, and I. S. Besedin, High fidelity two-qubit gates on fluxoniums using a tunable coupler, npj Quantum Inf. **8**, 130 (2022).

[352] A. Carignan-Dugas, S. K. Ranu, and P. Dreher, Estimating coherent contributions to the error profile using cycle error reconstruction, Quantum **8**, 1367 (2024).

[353] D. M. Debroy, E. Genois, J. A. Gross, W. Mruczkiewicz, K. Lee, S. Hong, Z. Chen, V. Smelyanskiy, and Z. Jiang, Context-aware fidelity estimation, Phys. Rev. Res. **5**, 043202 (2023).

[354] J. J. Wallman, M. Barnhill, and J. Emerson, Robust characterization of loss rates, Phys. Rev. Lett. **115**, 060501 (2015).

[355] T. Chasseur and F. K. Wilhelm, Complete randomized benchmarking protocol accounting for leakage errors, Phys. Rev. A **92**, 042333 (2015).

[356] C. C. López, A. Bendersky, J. P. Paz, and D. G. Cory, Progress toward scalable tomography of quantum maps using twirling-based methods and information hierarchies, Phys. Rev. A **81**, 062113 (2010).

[357] G. Tóth, W. Wieczorek, D. Gross, R. Krischek, C. Schwemmer, and H. Weinfurter, Permutationally invariant quantum tomography, Phys. Rev. Lett. **105**, 250403 (2010).

[358] A. Bendersky and J. P. Paz, Selective and efficient quantum state tomography and its application to quantum process tomography, Phys. Rev. A **87**, 012122 (2013).

[359] C. Greganti, M.-C. Roehsner, S. Barz, M. Waegell, and P. Walther, Practical and efficient experimental characterization of multiqubit stabilizer states, Phys. Rev. A **91**, 022325 (2015).

[360] A. Steffens, P. Rebentrost, I. Marvian, J. Eisert, and S. Lloyd, An efficient quantum algorithm for spectral estimation, New J. Phys. **19**, 033005 (2017).

[361] C. Carmeli, T. Heinosaari, J. Schultz, and A. Toigo, Probing quantum state space: Does one have to learn everything to learn something? Proc. Math. Phys. Eng. Sci. **473**, 20160866 (2017).

[362] J. Helsen, F. Battistel, and B. M. Terhal, Spectral quantum tomography, npj Quantum Inf. **5**, 74 (2019).

[363] J. Helsen, M. Ioannou, J. Kitzinger, E. Onorati, A. Werner, J. Eisert, and I. Roth, Shadow estimation of gate set properties from random sequences, Nat. Commun. **14**, 5039 (2023).

[364] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, The randomized measurement toolbox, Nat. Rev. Phys. **5**, 9 (2023).

[365] S. Aaronson, in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2018), p. 325.

[366] J. Kunjummen, M. C. Tran, D. Carney, and J. M. Taylor, Shadow process tomography of quantum channels, Phys. Rev. A **107**, 042403 (2023).

[367] S. T. Flammia and Y.-K. Liu, Direct fidelity estimation from few Pauli measurements, Phys. Rev. Lett. **106**, 230501 (2011).

[368] M. P. da Silva, O. Landon-Cardinal, and D. Poulin, Practical characterization of quantum devices without tomography, Phys. Rev. Lett. **107**, 210404 (2011).

[369] G. Gutoski and N. Johnston, Process tomography for unitary quantum channels, J. Math. Phys. **55**, 032201 (2014).

[370] X. Ma, T. Jackson, H. Zhou, J. Chen, D. Lu, M. D. Mazurek, K. A. G. Fisher, X. Peng, D. Kribs, K. J. Resch, *et al.*, Pure-state tomography with the expectation value of Pauli operators, Phys. Rev. A **93**, 032140 (2016).

[371] T. Moroder, P. Hyllus, G. Tóth, C. Schwemmer, A. Niggebaum, S. Gaile, O. Gühne, and H. Weinfurter, Permutationally invariant state reconstruction, New J. Phys. **14**, 105001 (2012).

[372] C. Schwemmer, G. Tóth, A. Niggebaum, T. Moroder, D. Gross, O. Gühne, and H. Weinfurter, Experimental comparison of efficient tomography schemes for a six-qubit state, Phys. Rev. Lett. **113**, 040503 (2014).

[373] M. Guţă, T. Kypraios, and I. Dryden, Rank-based model selection for multiple ions quantum tomography, New J. Phys. **14**, 105002 (2012).

[374] O. Landon-Cardinal and D. Poulin, Practical learning method for multi-scale entangled states, New J. Phys. **14**, 085004 (2012).

[375] T. Baumgratz, D. Gross, M. Cramer, and M. B. Plenio, Scalable reconstruction of density matrices, Phys. Rev. Lett. **111**, 020401 (2013).

[376] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Efficient quantum state tomography, Nat. Commun. **1**, 149 (2010).

[377] S. T. Flammia, in *17th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2022)*, Leibniz International Proceedings in Informatics (LIPIcs), Vol. 232, edited by F. Le Gall and T. Morimae (Schloss Dagstuhl—Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2022), p. 4:1.

[378] E. Van Den Berg, Z. K. Minev, A. Kandala, and K. Temme, Probabilistic error cancellation with sparse Pauli–Lindblad models on noisy quantum processors, Nat. Phys. **19**, 1116 (2023).

[379] E. van den Berg and P. Wocjan, Techniques for learning sparse Pauli-Lindblad noise models, Quantum **8**, 1556 (2024).

[380] M. Kieferová and N. Wiebe, Tomography and generative training with quantum Boltzmann machines, Phys. Rev. A **96**, 062327 (2017).

[381] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nat. Phys. **14**, 447 (2018).

[382] J. Gao, L.-F. Qiao, Z.-Q. Jiao, Y.-C. Ma, C.-Q. Hu, R.-J. Ren, A.-L. Yang, H. Tang, M.-H. Yung, and X.-M. Jin, Experimental machine learning of quantum states, Phys. Rev. Lett. **120**, 240501 (2018).

[383] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Reconstructing quantum states with generative models, Nat. Mac. Intell. **1**, 155 (2019).

[384] V. Gebhart, R. Santagati, A. A. Gentile, E. M. Gauger, D. Craig, N. Ares, L. Banchi, F. Marquardt, L. Pezzè, and C. Bonato, Learning quantum systems, Nat. Rev. Phys. **5**, 141 (2023).

[385] A. Fedorov, L. Steffen, M. Baur, M. P. da Silva, and A. Wallraff, Implementation of a Toffoli gate with superconducting circuits, Nature **481**, 170 (2011).

[386] J. Chu, X. He, Y. Zhou, J. Yuan, L. Zhang, Q. Guo, Y. Hai, Z. Han, C.-K. Hu, and W. Huang, Scalable algorithm simplification using quantum and logic, Nat. Phys. **19**, 126 (2023).

[387] In a doubly stochastic matrix, both the rows and columns sum to 1.

[388] H. F. Hofmann, Complementary classical fidelities as an efficient criterion for the evaluation of experimentally realized quantum operations, Phys. Rev. Lett. **94**, 160504 (2005).

[389] C. Figgatt, D. Maslov, K. A. Landsman, N. M. Linke, S. Debnath, and C. Monroe, Complete 3-qubit grover search on a programmable quantum computer, Nat. Commun. **8**, 1918 (2017).

[390] H. Levine, A. Keesling, G. Semeghini, A. Omran, T. T. Wang, S. Ebadi, H. Bernien, M. Greiner, V. Vuletić, H. Pichler, and M. D. Lukin, Parallel implementation of high-fidelity multiqubit gates with neutral atoms, Phys. Rev. Lett. **123**, 170503 (2019).

[391] C. Fang, Y. Wang, K. Sun, and J. Kim, Realization of scalable Cirac-Zoller multi-qubit gates, arXiv:2301.07564.

[392] Y. Lu, J. Y. Sim, J. Suzuki, B.-G. Englert, and H. K. Ng, Direct estimation of minimum gate fidelity, Phys. Rev. A **102**, 022410 (2020).

[393] X. Zhang, M. Luo, Z. Wen, Q. Feng, S. Pang, W. Luo, and X. Zhou, Direct fidelity estimation of quantum states using machine learning, Phys. Rev. Lett. **127**, 130503 (2021).

[394] B. M. Terhal, Quantum error correction for quantum memories, Rev. Mod. Phys. **87**, 307 (2015).

[395] M. A. Graydon, J. Skanes-Norman, and J. J. Wallman, Designing stochastic channels, arXiv:2201.07156.

[396] S. T. Flammia and J. J. Wallman, Efficient estimation of Pauli channels, ACM Trans. Quantum Comput. **1**, 1 (2020).

[397] Y. Chen, Z. Yu, C. Zhu, and X. Wang, Efficient information recovery from Pauli noise via classical shadow, arXiv:2305.04148.

[398] A. Carignan-Dugas, D. Dahlen, I. Hincks, E. Ospadov, S. J. Beale, S. Ferracin, J. Skanes-Norman, J. Emerson, and J. J. Wallman, The error reconstruction and compiled calibration of quantum computing cycles, arXiv:2303.17714.

[399] S. Chen, Y. Liu, M. Otten, A. Seif, B. Fefferman, and L. Jiang, The learnability of Pauli noise, Nat. Commun. **14**, 52 (2023).

[400] Here, a *k-body* error is any weight-*n* Pauli error acting on *k* gates. For example, a *Z* error is a weight-1 error acting on a single qubit, but both *IZ* and *ZZ* are (weight-1 and weight-2, respectively) *single*-body errors acting on two qubits involved in an entangling gate.

[401] E. Pelaez Cisneros, V. Omole, P. Gokhale, R. Rines, K. N. Smith, M. A. Perlin, and A. Hashim, Average circuit eigenvalue sampling on NISQ devices, arXiv:2403.12857.

[402] E. T. Hockings, A. C. Doherty, and R. Harper, Scalable noise characterisation of syndrome-extraction circuits with averaged circuit eigenvalue sampling, PRX Quantum **6**, 010334 (2025).

[403] U. Mahadev, in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2018), p. 259.

[404] Z. Brakerski, P. Christiano, U. Mahadev, U. Vazirani, and T. Vidick, A cryptographic test of quantumness and certifiable randomness from a single quantum device, J. ACM **68**, 1 (2021).

[405] G. D. Kahanamoku-Meyer, S. Choi, U. V. Vazirani, and N. Y. Yao, Classically verifiable quantum advantage from a computational bell test, Nat. Phys. **18**, 918 (2022).

[406] D. Zhu, G. D. Kahanamoku-Meyer, L. Lewis, C. Noel, O. Katz, B. Harraz, Q. Wang, A. Risinger, L. Feng, D. Biswas, *et al.*, Interactive protocols for classically-verifiable quantum advantage, arXiv:2112.05156.

[407] S. Ferracin, T. Kapourniotis, and A. Datta, Reducing resources for verification of quantum computations, Phys. Rev. A **98**, 022323 (2018).

[408] S. Ferracin, T. Kapourniotis, and A. Datta, Accrediting outputs of noisy intermediate-scale quantum computing devices, New J. Phys. **21**, 113038 (2019).

[409] S. Ferracin, S. Merkel, D. McKay, and A. Datta, Experimental accreditation of outputs of noisy quantum computers, Phys. Rev. A **104**, 042603 (2020).

[410] R. Blume-Kohout and K. C. Young, A volumetric framework for quantum computer benchmarks, Quantum **4**, 362 (2020).

[411] J. Hines and T. Proctor, Scalable full-stack benchmarks for quantum computers, IEEE Trans. Quantum Eng. **5**, 1 (2024).

[412] Metriq—community-driven quantum benchmarks, https://metriq.info/, 2024 (accessed: January 30, 2024).

[413] T. Tomesh, P. Gokhale, V. Omole, G. S. Ravi, K. N. Smith, J. Viszlai, X.-C. Wu, N. Hardavellas, M. R. Martonosi, and F. T. Chong, in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, 2022), p. 587.

[414] A. Li, S. Stein, S. Krishnamoorthy, and J. Ang, Qasmbench: A low-level quantum benchmark suite for NISQ evaluation and simulation, ACM Trans. Quantum Comput. **4**, 26 (2023) (2022).

[415] T. Lubinski, S. Johri, P. Varosy, J. Coleman, L. Zhao, J. Necaise, C. H. Baldwin, K. Mayer, and T. Proctor, Application-oriented performance benchmarks for quantum computing, IEEE Trans. Quantum Eng. **4**, 44 (2023).

[416] T. Lubinski, C. Coffrin, C. McGeoch, P. Sathe, J. Apanavicius, and D. E. B. Neira, Optimization applications as quantum performance benchmarks, ACM Trans. Quantum Comput. **5**, 18 (2024).

[417] N. P. Sawaya, D. Marti-Dafcik, Y. Ho, D. P. Tabor, D. E. B. Neira, A. B. Magann, S. Premaratne, P. Dubey, A. Matsuura, N. Bishop, *et al.*, Hamlib: A library of

Hamiltonians for benchmarking quantum algorithms and hardware, Quantum **8**, 1559 (2024).
[418] A. Miessen, D. J. Egger, I. Tavernelli, and G. Mazzola, Benchmarking digital quantum simulations above hundreds of qubits using quantum critical dynamics, PRX Quantum **5**, 040320 (2024).
[419] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).
[420] P. Murali, N. M. Linke, M. Martonosi, A. J. Abhari, N. H. Nguyen, and C. H. Alderete, in *Proceedings of the 46th International Symposium on Computer Architecture* (Association for Computing Machinery, New York, NY, USA, 2019), p. 527.
[421] P. Pfeuty, The one-dimensional Ising model with a transverse field, Ann. Phys. **57**, 79 (1970).
[422] D. M. Greenberger, M. A. Horne, and A. Zeilinger, in *Bell's Theorem, Quantum Theory and Conceptions of the Universe*, Vol. 69 (Springer Dordrecht, 1989), https://link.springer.com/book/10.1007/978-94-017-0849-4.
[423] While we model our errors using a *postgate* error matrix, it is equally valid to model errors using a *pregate* error matrix, or in some cases one which occurs concurrently with the gate [198].
[424] G. Lindblad, On the generators of quantum dynamical semigroups, Commun. Math. Phys. **48**, 119 (1976).
[425] D. Gottesman, Theory of fault-tolerant quantum computation, Phys. Rev. A **57**, 127 (1998).
[426] G. E. Crooks, Gates states and circuits (2020), https://threeplusone.com/pubs/on_gates.pdf.
[427] M. Pozniak, K. Zyczkowski, and M. Kus, Composed ensembles of random unitary matrices, J. Phys. A: Math. Gen. **31**, 1059 (1998).
[428] E. Bannai and E. Bannai, A survey on spherical designs and algebraic combinatorics on spheres, Eur. J. Comb. **30**, 1392 (2009).
[429] The Clifford group also forms a unitary 3-design only when the qudit dimension $D = 2$ [454–456].
[430] D. Gross, K. Audenaert, and J. Eisert, Evenly distributed unitaries: On the structure of unitary designs, J. Math. Phys. **48**, 052104 (2007).
[431] N. Goss, S. Ferracin, A. Hashim, A. Carignan-Dugas, J. M. Kreikebaum, R. K. Naik, D. I. Santiago, and I. Siddiqi, Extending the computational reach of a superconducting qutrit processor, npj Quantum Inf. **10**, 101 (2024).
[432] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. **58**, 13 (1963).
[433] N. Goss, A. Morvan, B. Marinelli, B. K. Mitchell, L. B. Nguyen, R. K. Naik, L. Chen, C. Jünger, J. M. Kreikebaum, D. I. Santiago, *et al.*, High-fidelity qutrit entangling gates for superconducting circuits, Nat. Commun. **13**, 7481 (2022).
[434] P. Liu, R. Wang, J.-N. Zhang, Y. Zhang, X. Cai, H. Xu, Z. Li, J. Han, X. Li, G. Xue, W. Liu, L. You, Y. Jin, and H. Yu, Performing SU(d) operations and rudimentary algorithms in a superconducting transmon qudit for $d = 3$ and $d = 4$, Phys. Rev. X **13**, 021028 (2023).
[435] S. Cao, M. Bakr, G. Campanaro, S. D. Fasciati, J. Wills, D. Lall, B. Shteynas, V. Chidambaram, I. Rungger, and P. Leek, Emulating two qubits with a four-level transmon qudit for variational quantum algorithms, Quantum Sci. Technol. **9**, 035003 (2024).
[436] M. Ringbauer, M. Meth, L. Postler, R. Stricker, R. Blatt, P. Schindler, and T. Monz, A universal qudit quantum processor with trapped ions, Nat. Phys. **18**, 1053 (2022).
[437] P. Hrmo, B. Wilhelm, L. Gerster, M. W. van Mourik, M. Huber, R. Blatt, P. Schindler, T. Monz, and M. Ringbauer, Native qudit entanglement in a trapped ion quantum processor, Nat. Commun. **14**, 2242 (2023).
[438] B. P. Lanyon, T. J. Weinhold, N. K. Langford, J. L. O'Brien, K. J. Resch, A. Gilchrist, and A. G. White, Manipulating biphotonic qutrits, Phys. Rev. Lett. **100**, 060504 (2008).
[439] Y. Chi, *et al.*, A programmable qudit-based quantum processor, Nat. Commun. **13**, 1166 (2022).
[440] G. Duclos-Cianci and D. Poulin, Kitaev's F_d-code threshold estimates, Phys. Rev. A **87**, 062338 (2013).
[441] H. Anwar, B. J. Brown, E. T. Campbell, and D. E. Browne, Fast decoders for qudit topological codes, New J. Phys. **16**, 063038 (2014).
[442] S. Muralidharan, C.-L. Zou, L. Li, J. Wen, and L. Jiang, Overcoming erasure errors with multilevel systems, New J. Phys. **19**, 013026 (2017).
[443] E. T. Campbell, H. Anwar, and D. E. Browne, Magic-state distillation in all prime dimensions using quantum Reed-Muller codes, Phys. Rev. X **2**, 041021 (2012).
[444] P. Gokhale, J. M. Baker, C. Duckering, N. C. Brown, K. R. Brown, and F. T. Chong, in *Proceedings of the 46th International Symposium on Computer Architecture*, ISCA '19 (Association for Computing Machinery, New York, NY, USA, 2019), p. 554.
[445] E. Gustafson, Noise improvements in quantum simulations of sqed using qutrits, arXiv:2201.04546.
[446] L. A. Truflandier, R. M. Dianzinga, and D. R. Bowler, Communication: Generalized canonical purification for density matrix minimization, J. Chem. Phys. **144**, 091102 (2016).
[447] S. Cao, D. Lall, M. Bakr, G. Campanaro, S. D. Fasciati, J. Wills, V. Chidambaram, B. Shteynas, I. Rungger, and P. J. Leek, Efficient characterization of qudit logical gates with gate set tomography using an error-free virtual Z gate model, Phys. Rev. Lett. **133**, 120802 (2024).
[448] L. M. Seifert, Z. Li, T. Roy, D. I. Schuster, F. T. Chong, and J. M. Baker, Exploring ququart computation on a transmon using optimal control, Phys. Rev. A **108**, 062609 (2023).
[449] B. Bollobás, *Modern Graph Theory*, Vol. 184 (Springer New York, NY, 1998).
[450] S. Chen, Z. Zhang, L. Jiang, and S. T. Flammia, Efficient self-consistent learning of gate set Pauli noise, arXiv:2410.03906.
[451] Z. Zhang, S. Chen, Y. Liu, and L. Jiang, Generalized cycle benchmarking algorithm for characterizing mid-circuit measurements, PRX Quantum **6**, 010310 (2025).
[452] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, Phys. Rev. X **8**, 031027 (2018).

[453] S. Ferracin, A. Hashim, J.-L. Ville, R. Naik, A. Carignan-Dugas, H. Qassim, A. Morvan, D. I. Santiago, I. Siddiqi, and J. J. Wallman, Efficiently improving the performance of noisy quantum computers, Quantum **8**, 1410 (2024).

[454] Z. Webb, The Clifford group forms a unitary 3-design, arXiv:1510.02769.

[455] H. Zhu, Multiqubit Clifford groups are unitary 3-designs, Phys. Rev. J. **96**, 062336 (2017).

[456] M. A. Graydon, J. Skanes-Norman, J. J. Wallman, Clifford groups are not always 2-designs, arXiv:2108.04200.