

Self-Testing Quantum Randomness Expansion on an Integrated Photonic Chip

Gong Zhang^{1,2,*} Ignatius William Primaatmaja^{1,3} Yue Chen¹ Si Qi Ng^{1,4} Hong Jie Ng^{1,3}
Marco Pistoia,^{5,6,7} Xiao Gong,¹ Koon Tong Goh,³ Chao Wang,¹ and Charles Lim^{1,†}

¹Department of Electrical & Computer Engineering, [National University of Singapore](#), Singapore

²State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang Key Laboratory of Optoelectronic Information Technology, College of Optical Science and Engineering, [Zhejiang University](#), Hangzhou, China

³Squareroot8 Technologies Pte Ltd, Singapore

⁴[Centre for Quantum Technologies](#), National University of Singapore, Singapore

⁵Global Technology Applied Research, [JPMorganChase](#), New York, New York 10017, USA

⁶IonQ, College Park, Maryland 20740, USA

⁷[IonQ Italia](#), Rome, RM, Italy



(Received 14 September 2025; revised 28 March 2026; accepted 21 April 2026; published 5 June 2026)

The power of quantum random number generation is more than just the ability to create truly random numbers—it can also enable *self-testing*, which allows the user to verify the implementation integrity of critical quantum components with minimal assumptions. In this work, we develop and implement a self-testing quantum random number generator (QRNG) chipset capable of generating 15.33 Mbits of certifiable randomness in each run, producing an expansion rate of 5.11×10^{-4} at a repetition rate of 10 MHz, with quantum side information considered for a high level of security. The chip design is based on a highly loss-and-noise-tolerant measurement-device-independent protocol, where random coherent states encoded using quadrature phase shift keying are used to self-test the integrated homodyne detectors—well-known to be challenging to characterize in practice. Importantly, this proposal opens up the possibility of implementing miniaturized self-testing QRNG devices at production scale using standard silicon photonics foundry platforms.

DOI: [10.1103/h7kt-m5fc](#)

I. INTRODUCTION

Random number generation (RNG) is important for a multitude of reasons, particularly for computing, cryptography, and gaming. Roughly speaking, the quality of random numbers can be evaluated using two metrics: uniformity and unpredictability [1]. The former looks at how much entropy the RNG device can produce—the higher the entropy, the more unbiased the output is. The latter studies the amount of residual entropy when the observer is given additional information about the device. In the ideal setting, the device should have maximum entropy even when the observer is given full information about the device.

The security of today's RNGs [2–5], be it classical or quantum in nature, depends critically on the

manufacturer's ability to characterize the behavior of the components. The characterization typically requires faithful theoretical modeling of the underlying physical process and the assumption that the device will not deviate from its expected behavior during operations. However, this approach may not be the most reliable for real-time entropy estimation, as all physical devices suffer from imperfections and degradations over time. Even though the output can pass statistical tests, but it has no protection against quantum side information, which means it provides only uniformity, not unpredictability.

A promising solution is to leverage *self-testing* QRNGs, which allow users to estimate real-time entropy values with minimal assumptions [6,7]. The basic idea is to leverage input-output measurement statistics that can only be explained by measurements on entangled quantum systems. This provides the user with a powerful tool to verify the use of entanglement in the generation process using purely classical methods. This, in turn, guarantees, statistically, that the random numbers have been generated in real time and contain sufficiently high entropy.

Across different self-testing QRNG protocols, a trade-off exists between security and practicality, with varying

*Contact author: zhanggong@zju.edu.cn

†Contact author: charles.lim@nus.edu.sg

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](#) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

assumptions about the devices. Device-independent (DI) QRNGs [8–13], which rely only on minimum physical assumptions and are independent of the physical implementation, face limitations during implementation due to the requirement of high-efficiency single-photon detectors and entangled photon sources, which restrict the system footprint and scalability.

Semi-DI-QRNGs improve feasibility by introducing practical assumptions, without compromising security. One such assumption is that the source states can be easily well-characterized while maintaining the independence of measurement devices, known as measurement-DI-QRNG (MDI-QRNG). Unlike source-DI-QRNGs, which require well-characterized detectors [14–17], MDI-QRNGs avoid common situations where a practical detector deviates from the ideal detector model due to unavoidable drift, imperfection, and adversarial manipulation. Additionally, the source side can be isolated from the environment easily and form a minimum and physically verifiable component.

Furthermore, the practicality of MDI-QRNG can be further improved by eliminating the need for single photon detection and designing protocols using room-temperature, high-speed, and high-efficiency homodyne detectors [18]. Additionally, to provide a complete security analysis, it is essential to assume that the adversary possesses quantum side information. Compared to other semi-DI protocols based on MDI assumptions [19–21] and bounded dimension or energy assumptions [22–25], an MDI-QRNG protocol that accounts for quantum side information and employs homodyne detectors is able to provide high levels of security and practicality. This approach is fully compatible with photonic integration, which is necessary for miniaturized electronic devices.

Most chip-based QRNG demonstrations remain in the trusted-device category, based on vacuum fluctuation [26,27], phase noise [28], and qubit states collapse [29], with only a few demonstrating chip-based source-DI [30–32] or energy-bound QRNGs [33]. Fully integrated chip-based MDI-QRNGs using homodyne detection have not yet been realized.

Various material platforms are available for chip-based QRNGs [34–37], including lithium niobate [34], group III-V metals [35], perovskite [38], and, notably, silicon [36, 37]. The silicon platform stands out due to its high-speed modulators and detectors, allowing for the integration of all necessary components to encode and decode quantum states, including laser and electronic circuits [39–41]. Recent demonstrations in quantum computing [42–44] and quantum communication [45–49] applications have underscored the capability of silicon photonic devices to encode and decode quantum states. However, silicon has inherent limitations, particularly in linearity and phase-loss dependency on high-speed modulators and the efficiency and accuracy of photodiodes, which necessitate careful design considerations. Although the quadrature phase shift keying

(QPSK) modulation and homodyne detection scheme on a silicon chip has been demonstrated in classical communication applications [50], the classical requirement is more on the speed and signal-to-noise ratio rather than the accuracy of prepared states. Thus, an innovative redesign of classical components is necessary.

In this work, we demonstrate the first self-testing QRNG chip with a fully integrated encoder and decoder, operating at room temperature using a homodyne detector. Our contributions to this work are twofold: (1) Theoretically, we develop a robust self-testing MDI-QRNG protocol that enhances the randomness expansion rate and tolerates more losses and noises on the chip. The protocol employs discrete modulated coherent states in QPSK, combined with homodyne detection. The modulation format is flexible and can be modified based on the requirement of speed and efficiency. With the quantum side-information assumptions, users can self-test randomness during operation by statistically verifying the detector’s response against an ideal detector. (2) Experimentally, we design a silicon photonic chip integrating all necessary photonic components except the laser. Using 8 in. MPW processes and components, the fabricated chips approach the production level. The modulation scheme is designed to address the limitation of phase-loss dependency. The basis selection phase modulator utilizes a Mach-Zehnder modulator in a push-push configuration to compensate for the loss differences. An IQ (in-phase and quadrature) modulator with small driving voltages minimizes the effect of phase-loss dependency. The chip-based homodyne detector achieves a total efficiency of 69.1%, and we obtain a secure randomness expansion rate of 5.11×10^{-4} at a 10 MHz repetition rate. In each run, the experiment generated 15.33 Mbits of fresh randomness.

This work combines theoretical and experimental schemes, demonstrating fully integrated silicon photonic chips for self-testing QRNG and providing a new level of security on the chip. Our chip paves the way toward integrating practical QRNGs in systems with a small footprint and is secure.

II. SELF-TESTING QUANTUM RANDOMNESS EXPANSION PROTOCOL

Our randomness expansion protocol is based on a prepare and measure (P&M) game $\mathcal{G}$. P&M game $\mathcal{G}$ is an analog of the more well-known non-local games; the latter has applications in Bell nonlocality [51,52] and device-independent protocols [53]. While nonlocal games are typically studied in entanglement-based scenarios where all the players are receivers of some quantum states, P&M games typically involve a sender and a receiver. The game $\mathcal{G}$ can be characterized using the quantum state encoding, the probability distribution of the inputs for the sender and the receiver, and a scoring rule.

In each P&M game round, Alice and Bob randomly choose between test and generation modes, where Alice prepares one of four trusted QPSK coherent states and Bob selects a corresponding homodyne measurement basis. Bob's continuous detector outputs are discretized and evaluated using a scoring rule [Fig. 1(b)] during test rounds to verify that the measurement device behaves consistently with the expected output statistics. If the observed score distribution passes the acceptance criteria, a quantum-proof extractor is applied to the raw data to produce the final certified random bits. Any nonideal behaviors, such as from phase-dependent loss or intensity imbalance, are bounded and explicitly incorporated into the entropy analysis via the acceptance criteria. More details are shown in the Methods section.

Note that our randomness certification does not assume that the homodyne detector and quantum channel are characterized, nor do we assume that they behave independently or identically for each round. This provides our protocol with a practical advantage because detectors are typically hard to model and fully characterize. As long as the score distribution in the actual implementation closely resembles the one that originates from a homodyne detection, the protocol can certify the randomness of the measurement outcomes. With fine-grained binning, the protocol can certify the measurement device better and tolerate more losses and noises.

Moreover, we allow the adversary to have some pre-shared entanglement with Bob's untrusted detector before the protocol starts, i.e., the adversary possesses quantum side information, but we assume that the adversary does not update her quantum side information during the protocol operation. This is required by the entropy accumulation theorem (EAT) used in our security proof, namely that the inputs and outputs satisfy the so-called Markov chain condition [54,55]. This assumption is not too restrictive because the quantum channel is not exposed to any external user during the protocol operation, and the whole chip is inside a secure location. Additionally, a fully integrated chip cannot be easily injected or probed with a signal through unexpected ports unless the chip is specifically designed to allow such access.

As a result, the protocol is self-testing for the user. They only need to trust the quantum state preparation, while the performance of detectors and the generated randomness can be certified via the parameter estimation step during the operation.

The security of our protocol is guaranteed by the quantum-proof strong extractor, such that either the protocol is aborted with high probability (e.g., in the case where the device does not behave as expected) or the output string is close to an ideal random string, which is uniformly distributed and independent from any quantum side information and especially from the initial random seeds. The final certified randomness arises exclusively

from the quantum prepare-and-measure statistics, not from the input seed. It is fundamentally different from simply using random seed as the random source.

Since our protocol consumes random bits during the input generation step, the net randomness expansion rate $r_{\text{net}} := (\ell - \ell_{\text{in}})/n$ is the appropriate figure-of-merit for the protocol rather than the randomness generation rate. Here, ℓ_{in} is the expected amount of randomness used during the protocol operation. Using the entropy accumulation theorem (EAT), our protocol has an expected randomness expansion rate of

$$r_{\text{net}} \gtrsim h - \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \quad (1)$$

where h depends on the tolerated score distribution and $\mathcal{O}(1/\sqrt{n})$ is a correction term for the finite-size effects, which scales with $1/\sqrt{n}$. The explicit expression for the correction term can be found in the Methods section (Sec. VIII B).

III. SILICON PHOTONIC CHIP DESIGN

A silicon photonic chip is designed to implement our randomness expansion protocol. The chip schematic and a microscopic photo are shown in Figs. 1(c) and 1(d), respectively. The chip is fabricated by Advanced Micro Foundry Singapore using their active silicon-on-insulator (SOI) platform. The operation wavelength is chosen at O-band (1310.4 nm). Although, at this stage, we use an external laser and grating coupled to the chip, the full integration of laser and other control and supporting electronics is possible.

The local oscillator (LO) path uses a high-speed carrier-depletion Mach-Zehnder modulator (MZM) with a specially designed driving scheme for phase selection between 0 and $\pi/2$. The two electrodes are supplied with two driving signals P_+ and P_- with different intensities to compensate for the phase-dependent loss caused by the silicon modulators. More detail is covered in Sec. IV. An additional thermo-optical phase shifter on the LO path compensates for all residual phase differences.

The signal path encodes the required QPSK states using an IQ modulator made of two parallel MZMs driven by push-pull signals. We choose a very small driving voltage of ± 0.1 V, at which the phase-loss dependency is negligible. The phase shifters inside the IQ modulator act as bias control, modifying the working point and ensuring the output QPSK states have the same intensity. The encoded states are then attenuated to the required intensity.

The signal and LO are finally connected to a homodyne detector with a tunable BS to fine-tune the detector balance. We use directional couplers rather than multimode interferometers (MMIs) for their lower insertion loss. The photodiode output currents i_1 and i_2 are subtracted directly

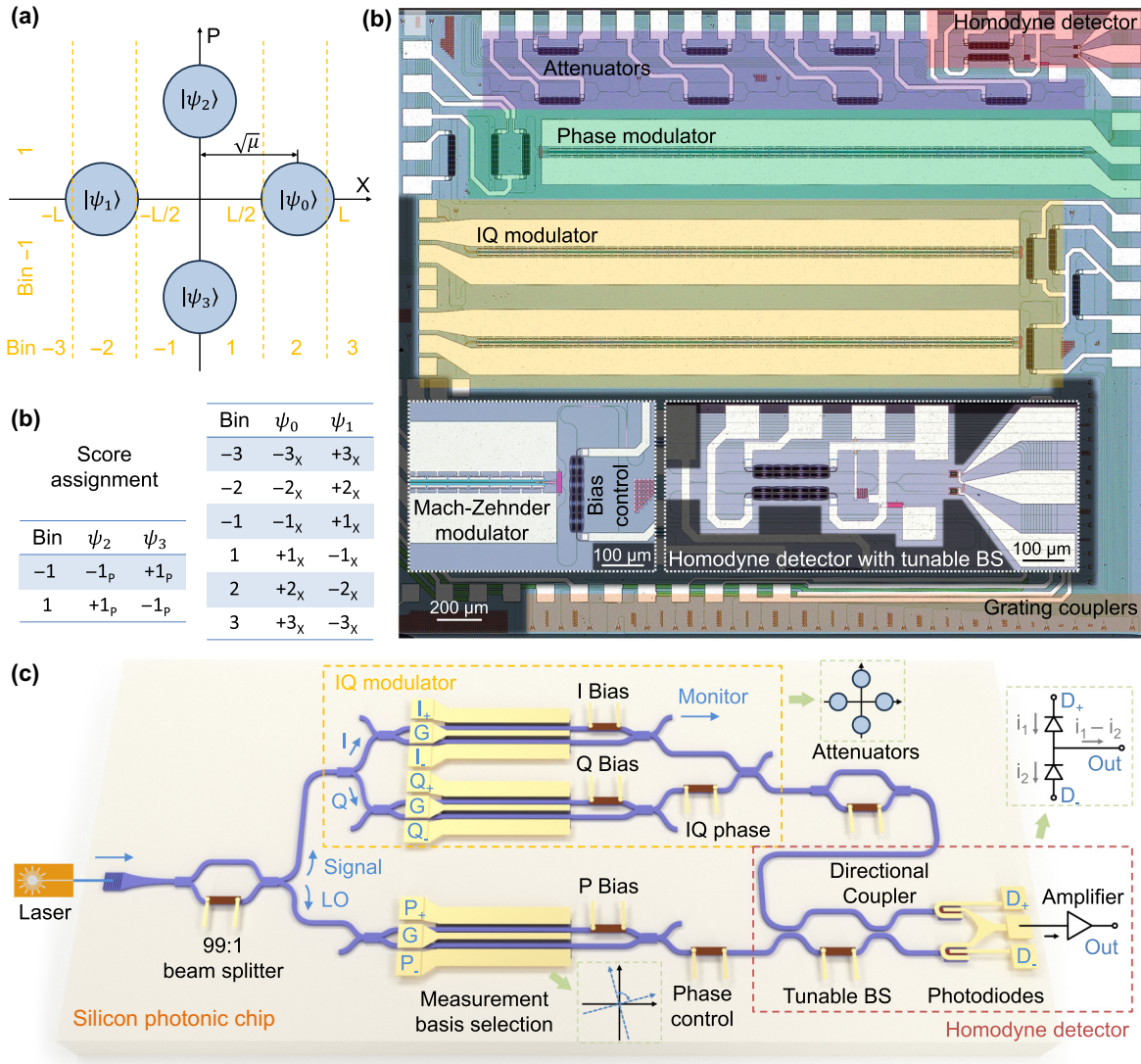


FIG. 1. Overview of the self-testing quantum randomness expansion chip. (a) Schematic of the encoding and decoding scheme. The input states are in quadrature phase shift keying with fixed intensity μ . The receiver measures either X or P quadrature and discretizes the output into six or two bins, respectively. (b) The score of each round is assigned depending on the input states and the measured output bins. (c) Schematic of the quantum randomness expansion chip design. The laser is off-chip coupled and then split to signal and local oscillator path using a Mach-Zehnder interferometer as a tunable 99:1 beam splitter. The local oscillator path uses a high-speed carrier-depletion Mach-Zehnder modulator with a specially designed driving scheme for measurement basis selection. Details can be found in Sec. IV. The signal path uses an IQ modulator based on an Mach-Zehnder interferometer structure for encoding, where each arm contains a high-speed Mach-Zehnder modulator and modulates one of the I or Q quadratures. To form the IQ modulation, the modulated signals from both arms are combined with a fixed phase, set by the phase shifter named IQ phase. The driving signals of each Mach-Zehnder interferometer are supplied with a push-pull configuration, which means that the center electrode is grounded, while the top and bottom signals satisfy $I_+ = -I_-$ and $Q_+ = -Q_-$. The bias controllers within Mach-Zehnder modulators modify the working point and ensure the output quadrature phase shift keying states have the same intensity. The port marked as Monitor is connected to an external power meter for fiber-chip alignment. The output from the IQ modulator is then connected to a series of attenuators made by a Mach-Zehnder interferometer. The signal and local oscillator are finally connected to the input ports of a homodyne detector. The homodyne detector uses a tunable 50:50 beam splitter made of directional couplers to mix the input signal and local oscillator. The beam splitter outputs to two identical high-efficiency germanium photodiodes. A common electrode connects the anode and the cathode of each photodiode, such that the common electrode outputs the subtracted photocurrent $i_1 - i_2$. The other electrodes D_+ and D_- provide proper reverse bias for the photodiodes. The substrate photocurrent passes through a customized trans-impedance amplifier for detection. (d) A microscopic photo of the fabricated chip with false-colored key components. The Mach-Zehnder modulator with its bias controller and the homodyne detector is zoomed in and shown in the inset.

on the chip. The substrate photocurrent passes through a customized transimpedance amplifier and is sampled using an oscilloscope. Detailed analysis for the homodyne detector is covered in Sec. V.

IV. PHASE-LOSS INDEPENDENT MODULATION

The plasma dispersion effect is the most common method to achieve high-speed modulation on silicon photonic platforms. These modulators have a natural dependency between the output phase and loss. In classical communication applications, such dependency has limited effect on the modulators' speed and signal-to-noise ratio and can be negligible when the driving voltage is low. Implementing a 0 to $\pi/2$ large modulation is unnecessary. The I/Q measurement is achieved using a 90-degree hybrid with a heterodyne detection scheme. However, in our application, this dependency on the encoder side can distort the prepared states. On the receiver side, the basis selection is required to avoid the 3-dB loss associated with heterodyne detection, and such large phase modulation could be associated with intensity change, which is a severe security loophole. In this section, the effect of phase-loss dependency is analyzed, and new solutions are introduced.

A schematic for the high-speed modulator is shown in Fig. 2(a). The phase-loss dependency is measured using a push-pull configuration, with $V_1 = -V_2$ and $\Phi_1 = -\Phi_2$. The input wavelength is scanned across the whole O-band to acquire the Mach-Zehnder interferometer (MZI) extinction ratio and phase, which can further derive the loss. The difference in loss and phase as a function of the applied voltage is shown in Fig. 2(b). The fast-increasing zone above ± 0.6 V corresponds to one of the diodes with forward bias. The data can further derive the linear relationship between the phase and loss, as shown in Fig. 2(c). A phase modulator can introduce more than 35% power difference in the 0– 1.5π range, which proves the necessity of using an IQ modulator instead of a phase modulator for QPSK encoding. For our IQ modulator, the ± 0.1 V driving voltage corresponds to a phase shift of 0.04π and a loss difference of maximum 0.05 dB, which can be compensated easily by the bias control.

The requirement for the phase modulator on the LO path differs from the IQ modulator. The phase modulator requires a phase range of 0 to $\pi/2$ with minimum insertion loss to maintain the high efficiency of the homodyne detector. Here, we propose to operate the MZM in a push-push configuration, with $\Phi_2 = r\Phi_1$, where the nonlinearity of the MZM response compensates for the phase-dependent loss. Figures 2(c) and 1(d) show the simulated output intensity and phase of MZM, which indicates that the output intensity fluctuates more and has a higher peak intensity when the ratio r decreases. Considering a fixed $\pi/2$ phase range, an optimum working point with $r = 0.6$

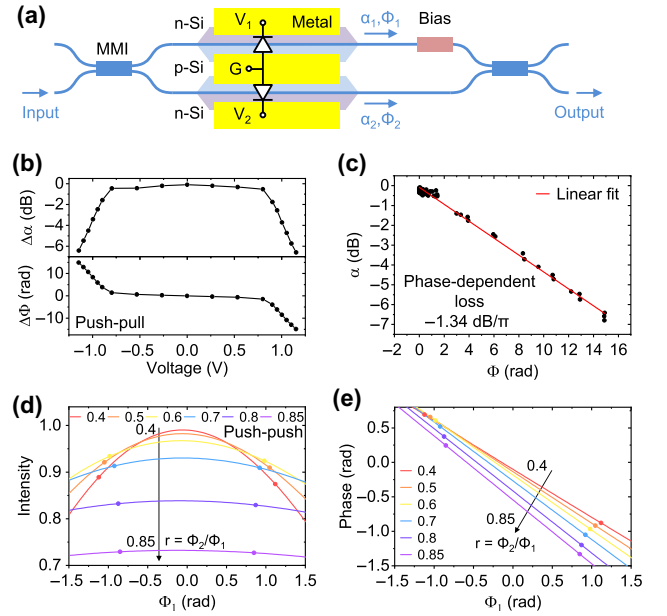


FIG. 2. Modulator phase-loss dependency analysis. (a) Schematic of the chip-based high-speed Mach-Zehnder modulator. The modulator is based on an Mach-Zehnder interferometer structure, with carrier depletion phase modulators on both arms and their polarity reversed. The center electrode is connected and grounded, while voltages V_1 and V_2 are applied on each arm. After the modulator section, the output light gains phase Φ and loss α . After an additional phase shifter for bias control, the lights are mixed on a multimode interferometer and output from one of the output ports. (b) The measured loss and phase difference between the two arms as a function of supplied voltages, when $V_1 = -V_2$. (c) Extracted phase-loss dependency from (b), showing a linear relationship. (d) and (e) Simulated output intensity (d) and phase (e) from a push-push Mach-Zehnder modulator under different modulated phases on each arm. All the curves use optimum bias values. The dots on each curve correspond to working points with phase ranges of $\pi/2$ and identical intensities.

exists, where the output light has no phase-dependent loss and minimum total loss.

V. HIGH-EFFICIENCY HOMODYNE DETECTOR

The homodyne detector is another key component of our protocol. Although the protocol does not require the detector to be characterized, we provide the specifications as a reference here to assist in demonstration.

The quantum efficiency of photodiodes is measured first. The current-voltage (I - V) response of the photodiode is shown in Figs. 3(a) and 3(b). An important feature to notice is the slight increase in photocurrent above 3 V, mainly due to impact ionization. Figure 3(c) clearly shows that the responsivity of the photodiode at 4 V is higher than at 1 V. However, in our application, quantum efficiency is defined as the effectiveness of converting incident photons into electrons. Because the impact ionization process happens

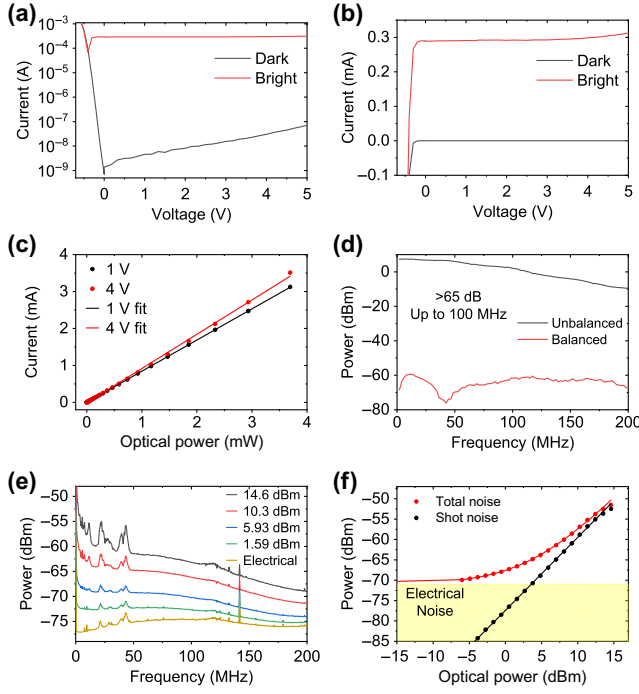


FIG. 3. Homodyne detector performance. (a) and (b) Current-voltage (I - V) curve of photodiodes in (a) logarithmic scale and (b) linear scale, in both dark and bright conditions. The dark count is more than four orders of magnitude lower than photocurrents below 4 V, essential to reduce any electronic noise in the homodyne detector. (c) The photocurrent as a function of input optical power at different reverse bias voltages. (d) The detector balance test shows a common mode rejection ratio of >65 dB up to 100 MHz. (e) Shot-noise spectrum at different local oscillator power, showing >11 dB shot noise to electronic noise clearance up to 100 MHz. (f) Shot noise as a function of local oscillator power at 10 MHz clock frequency. More than 19 dB maximum clearance and the saturation power at about 13 dBm local oscillator power are observed. The power scale is different due to the difference in resolution bandwidth.

after the photon-electron conversion process; thus, the increase in photocurrent and responsivity can only be considered a gain rather than a quantum efficiency increase. As a result, the quantum efficiency should be measured at a low reverse bias voltage of 1 V and the flat region on I - V curves, but the photodiodes can still be used at 4 V bias to provide a higher gain. We sample three different chips and get a quantum efficiency range of 76.9%–79.9%. The lower bound of 76.9% is used as the quantum efficiency value in our key rate calculation to maintain security.

The photocurrent subtraction is achieved directly on the chip. The common mode rejection ratio (CMRR) is fine-tuned using a tunable MZI. The measured CMRR is >65 dB within the 100 MHz 3 dB bandwidth, as shown in Fig. 3(d). The CMRR confirmed that the photodiode performances on the chip are similar, and the signal traces and wire bonds have a minimal effect on the signal integrity.

The measured shot-noise spectrum is shown in Fig. 3(e). Although the 3 dB bandwidth is about 100 MHz, the shot noise-limited measurement can reach more than 200 MHz. Limited by the bandwidth of the homodyne detector, we choose a system repetition rate of 10 MHz, while the sample rate for the signal source and oscilloscope is kept at 1 GS/s. The shot noise as a function of local oscillator power at 10 MHz is shown in Fig. 3(f). Considering the insertion loss of the phase modulator on the LO path, the shot-noise clearance in the experiment is 14.2 dB, which is lower than the saturation power, corresponding to a noise-equivalent efficiency $\eta_{\text{eq}} = 96.2\%$. Another loss we consider within the homodyne detector is the insertion loss of the tunable MZI structure based on directional couplers. The total insertion loss is 0.30 dB = 93.3%. As a reference, we also measured the insertion loss of a tunable MZI made of MMI as 0.72 dB = 84.7%, which is too high for our protocol. The total homodyne efficiency, considering the quantum efficiency of photodiodes, the insertion loss of the beamsplitter, and the noise equivalent efficiency, is $\eta = 69.1\%$.

Moreover, we customized a set of new photonic chip components that can potentially provide better performance. The new photodiode at 1550 nm wavelength has a quantum efficiency of 92.4%. Tunable beam splitter insertion loss is 0.30 dB = 93.3%. Assume it can reach the same shot-noise clearance $\eta_{\text{eq}} = 96.2\%$. The total homodyne efficiency can reach 83.0%.

VI. RANDOMNESS EXPANSION SIMULATION AND EXPERIMENT

With the chip performance well characterized, the experimental parameters related to the randomness expansion protocol are optimized and listed in Table I. Based on Eq. (1), the simulated $\mathbf{r}_{\text{net}}$ is shown in Fig. 4. Figure 4(a) shows the effect of homodyne detection efficiency and the number of discretized bins on the homodyne detector output. As expected, more bin numbers increase the randomness expansion rate and possess more tolerance for detector efficiency. The data for two bins shows a minimum acceptable efficiency of about 86%, while the result

TABLE I. Experiment parameters.

Homodyne efficiency η	69.1%
Number of bins $2m$	6
Number of rounds n	3×10^{10}
State amplitude $\sqrt{\mu}$	0.0672
Test probability γ	0.12
Completeness error ϵ_{com}	1×10^{-3}
Soundness error ϵ_{sou}	1×10^{-6}
Entropy accumulation error ϵ_{EA}	1×10^{-6}
Randomness extraction parameter ϵ_{ext}	1×10^{-6}
Smoothing parameter ϵ_s	4.99×10^{-7}

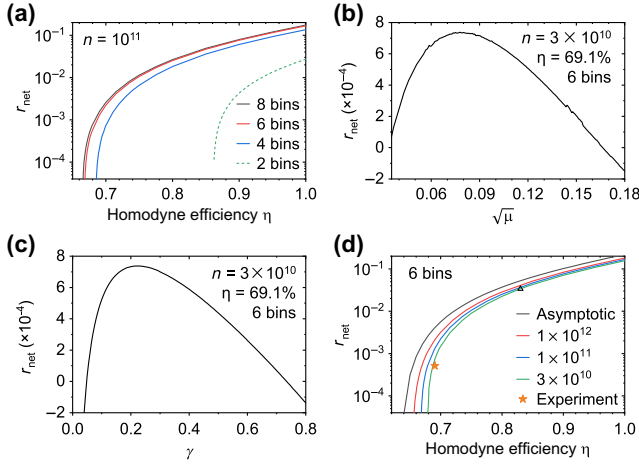


FIG. 4. Quantum randomness expansion rate simulation. (a) Rate as a function of homodyne detection efficiency with different numbers of discretization bins. A higher bin number shows better loss tolerance. The two bins result is reproduced from Ref. [18] (b) and (c). The effect of (b) input state amplitude and (c) test probability on the expansion rate. (d) Asymptotic rate and finite size rate at different numbers of rounds. The star marks our experimental result with a homodyne efficiency of 69.1% and rounds number of 3×10^{10} . The triangle marks the simulation point with a potentially upgraded homodyne efficiency of 83.0%, showing about two orders of magnitude improvement of the randomness expansion rate.

for four to eight bins shows a significant improvement at lower efficiency down to 67%. Such improvement is essential considering the relatively higher loss and noise on photonic chips.

Each point in Fig. 4(a) is optimized with respect to the state amplitude $\sqrt{\mu}$ and test probability γ , which are closely related to the experimental settings. Figures 4(b) and 4(c) show the effect of these two parameters on the r_{net} . At the efficiency of our chip $\eta = 69.1\%$, the input state amplitude that provides a positive rate can vary from 0.035 to 0.165, while the test probability can vary from 0.05 to 0.73. In our experiment, we tried to keep the state amplitude optimum but used a slightly lower test probability to increase the tolerated score range δ and reduce the required accuracy of the experiment. Figure 4(d) shows the finite size rate with six bins. Considering our chip-based homodyne efficiency of 69.1%, we choose the number of rounds $n = 3 \times 10^{10}$.

The measured quadrature distribution for all four test states is shown in Fig. 5, which faithfully resembles Gaussian distributions. Based on the scoring rule shown in Fig. 1(b), Table II shows the target score ω , tolerance δ , and experiment-measured error. Thus, the data is accepted and can proceed to the postprocessing step. The final secure random expansion rate is 5.11×10^{-4} bit/symbol. With the number of rounds $n = 3 \times 10^{10}$, the experiment generated 15.33 Mbits of fresh randomness in total. The

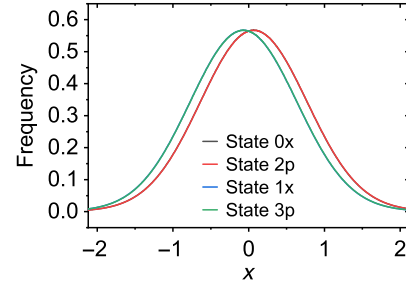


FIG. 5. Output quadrature distribution. The measured quadrature density distribution with all test states. The results clearly resemble Gaussian distributions.

actual randomness expansion rate is about 64 bps and is limited by our oscilloscope data transfer and system calibration time, see Method Sec. VIII C. We want to emphasize that the randomness rate is relatively low due to the assumption that the adversary possesses quantum side information, which is the highest security assumption.

VII. DISCUSSION

As a proof-of-principle experiment, the current required number of rounds, $n = 3 \times 10^{10}$, still has room for improvement, primarily constrained by homodyne efficiency. Our new customized high-efficiency waveguide photodiode in the lab achieves 92.4% efficiency at 1550 nm wavelength, providing a baseline for further enhancements. Shot-noise clearance can be improved through increased local oscillator power and improved amplifiers.

Homodyne measurements offer the advantage of reaching higher system bandwidth at room temperature compared to single-photon detectors. Considering chip integration, the active components, including modulators and detectors, are designed for classical communication with a bandwidth exceeding 20 GHz. The homodyne detector with a bandwidth of up to 20 GHz has also been demonstrated by our group [56,57], providing a substantial increase in the system clock rate by 200 times. Taking the detector efficiency increase into account, the randomness expansion rate per round reaches 0.0338. We anticipate the randomness expansion rate can reach $2 \text{ GS/s} \times 0.0338 = 68 \text{ Mbps}$ with current available technologies. The practical

TABLE II. Experimental score distribution.

Score	ω	δ	$f(c)/\gamma - \omega$
-3_X	0.0659	1.55×10^{-5}	0.61×10^{-5}
-2_X	0.0688	1.58×10^{-5}	-1.45×10^{-5}
-1_X	0.0931	1.83×10^{-5}	0.11×10^{-5}
-1_P	0.2278	2.85×10^{-5}	-0.79×10^{-5}
$+3_X$	0.0927	1.83×10^{-5}	-0.36×10^{-5}
$+2_X$	0.0811	1.71×10^{-5}	1.43×10^{-5}
1-all	0.3706	3.60×10^{-5}	0.46×10^{-5}

realization of this full upgrade remains important future work.

A missing component for full chip integration is the on-chip laser, but it has already been demonstrated in major photonic foundries using monolithic, heterogeneous, and hybrid integration methods [58–60]. The key parameters, such as a maximum power of 100 mW at 1550 nm and a linewidth below 1 MHz, are aligned closely with the table-top lasers we use now, indicating anticipated performances similar to those of this manuscript. The self-testing nature of our protocol provides an additional layer of verification that detects any laser-induced deviations.

Real-time hashing can also be implemented to improve the completeness of the system. A two-universal hash function, like Toeplitz-matrix hashing, with sampled sub-block hashing method can be efficiently implementable on field-programmable gate array (FPGA) or application-specific integrated circuit (ASICs) without excessive resources [61].

Our current protocol involves a trade-off between system complexity and the randomness expansion rate. Similar to our recent work [18], increasing the number of bins and input states not only increases the number of statistics that need to be calculated and consumes more input randomness, but also imposes higher demands on component performance, such as the linearity of IQ modulator and the accuracy of homodyne detectors. Users can select the appropriate configuration based on their specific application requirements.

In conclusion, we propose and demonstrate a self-testing QRNG using a silicon photonic chip. The upgraded protocol, incorporating an uncharacterized measurement device, can tolerate detector efficiency down to 67%. The silicon photonic chip integrates all necessary components, except the laser, on the same platform. The encoding uses an IQ modulator and a Mach-Zehnder modulator in a push-push configuration for detector basis selection. Both modulators eliminate the effect of phase-loss dependency. The chip-based homodyne detector achieves a high total efficiency of 69.1% at 1310 nm wavelength. The QRNG chip provides a total randomness expansion rate of 5.11×10^{-4} at a repetition rate of 10 MHz.

A comparison with state-of-the-art QRNG chips is shown in Table III. Our MDI-QRNG achieves the highest security level demonstrated on a chip to date. It is the first demonstration of MDI-level security with quantum side-information protection and is compatible with full chip integration. The modest 64 bps rate must be viewed in this context. It is worth clarifying that the practical security of a QRNG can be quantified by the number and strength of trusted assumptions required by each security model. The security hierarchy is generally DI (trust nothing) > semi-DI (trust something) > trusted device (trust all), while the achievable bitrate follows the opposite trend, DI < semi-DI < trusted device. Moreover, for MDI-QRNG, it uses much simpler equipment but delivers high security with only a minimally trusted source, while the entire detection chain remains untrusted. This is particularly important since detectors are the most attack-prone components. Our work paves the way for the application of quantum random numbers, providing provable security across diverse fields such as artificial intelligence, healthcare, energy services, blockchain, internet of things, and financial applications.

VIII. METHODS

A. Self-testing quantum randomness expansion protocol

Our randomness expansion protocol is based on a prepare and measure (P&M) game $\mathcal{G}$. P&M game $\mathcal{G}$ is an analog of the more well-known nonlocal games; the latter has applications in Bell nonlocality [51,52] and device-independent protocols [53]. While nonlocal games are typically studied in entanglement-based scenarios where all the players are receivers of some quantum states, P&M games typically involve a sender and a receiver. The game $\mathcal{G}$ can be characterized using the quantum state encoding, the probability distribution of the inputs for the sender and the receiver, and a scoring rule.

A scoring rule assigns a score to each input-output configuration. In our proposed protocol, the score distribution is monitored to certify the amount of randomness generated. The scoring rule in this work can be seen in Eq. (2).

TABLE III. Comparison with state-of-the-art quantum random number generator chips.

Work	Security model	Side information	Level of integration	Certified entropy rate
[26]	Trusted device	Quantum	Chip PD and TIA	100 Gbit/s
[27]	Trusted device	Quantum	Chip PD and TIA	18.8 Gbit/s
[28]	Trusted device	Classical	Laser and PD chip	8 Gbit/s
[29]	Trusted device	Classical	SPAD chip	310 Mbit/s
[30]	Source-DI	Quantum	Heterodyne receiver chip	20.0 Gbit/s
[31]	Source-DI	Quantum	Heterodyne receiver chip	13.4 Gbit/s
[32]	Source-DI	Quantum	Decoder chip with detector off-chip	9.49 Mbit/s
[33]	MDI	Classical	SPAD source and detector	6 kbit/s
Our	MDI	Quantum	Fully integrated encoder and detector	64 bit/s (Test) 68 Mbit/s (Simulation)

Our proposed quantum random number generation protocol can be found below

- (1) *Input generation.* For each round $i \in [n]$, the round is randomly chosen between test round ($T_i = 1$) and generation round ($T_i = 0$) with test probability γ and key probability $1 - \gamma$. In each test round, the encoder, named Alice, selects her input as $X_i \in \mathcal{X} = \{0, 1, 2, 3\}$, with an equal probability of 0.25 for each input. The receiver, named Bob, selects his input as $Y_i = 0$ when $X_i \in \{0, 1\}$ and $Y_i = 1$ when $X_i \in \{2, 3\}$. In each generation round, Alice and Bob fix their input as $X_i = 0$ and $Y_i = 1$.
- (2) *Quantum state preparation.* Alice prepares trusted quantum states $|\psi_{X_i}\rangle \in \{|\sqrt{\mu}e^{ix\pi/2}\rangle : x \in \mathcal{X}\}$ in the form of QPSK with fixed intensity μ based on her input X_i , as shown in Fig. 1(a).
- (3) *Quantum state measurement.* Bob measures the states prepared by Alice using an untrusted homodyne detector with his setting $Y_i \in \{0, 1\}$, which corresponds to the local oscillator phase of 0 and $\pi/2$. The continuous output is discretized into $2m$ bins $\mathcal{B} = \{\pm 1, \pm 2, \dots, \pm m\}$ by dividing a finite interval $(-L, L)$ into equal $(2m - 2)$ intervals, and the discretized result is recorded as B_i . In our current experiment, the number of bins is chosen as 6 and 2 when $Y = 0$ and 1, respectively.
- (4) *Parameter estimation.* For all generation rounds ($T_i = 0$), we set $C_i = \perp$. On the other hand, for the test rounds ($T_i = 1$), the scores $C_i = c$ are determined based on the following scoring rule:

$$c = \begin{cases} +b_X, & (X = 0, B = +b \text{ or } X = 1, B = -b) \\ -b_X, & (X = 0, B = -b \text{ or } X = 1, B = +b) \\ +b_P, & (X = 2, B = +b \text{ or } X = 3, B = -b) \\ -b_P, & (X = 2, B = -b \text{ or } X = 3, B = +b) \end{cases}. \quad (2)$$

The score assignment in our current experiment is listed in Fig. 1(b). The frequencies $f(c)$ of each score c from all n rounds are calculated as

$$f(c) = \frac{1}{n} \sum_{i=1}^n \delta_{C_i, c}, \quad (3)$$

where $\delta_{j,k}$ denotes the Kronecker delta. The protocol is accepted if

$$f(c) \leq \gamma(\omega_c + \delta_c), \quad \forall c \neq \perp, \quad (4)$$

where ω is the ideal test score distribution, and δ is the tolerated score range, which means all the frequencies for $c \neq \perp$ should fall within a tolerated score distribution. Otherwise, the protocol is aborted.

- (5) *Randomness extraction.* Our raw random string is formed by concatenating $\mathbf{R} = (\mathbf{B}, \mathbf{T}, \mathbf{X}, \mathbf{Y})$. A quantum-proof strong extractor Ext with some uniformly random seed $\mathbf{S}$ is applied to the raw random string and outputs a ℓ -bit string $\mathbf{Z} = \text{Ext}(\mathbf{R}, \mathbf{S})$. The protocol finally outputs the concatenated string $\mathbf{K} = (\mathbf{Z}, \mathbf{S})$.

Note that the input randomness is recycled by hashing the random inputs together with the measurement outputs, which allows the protocol to consume more input randomness while maintaining a positive net random expansion rate. To be able to recycle the input randomness, the inputs are assumed to be isolated and are not leaked during and after the protocol. Also, the input is chosen randomly from a private and trusted source of randomness in an independent and identically distributed (i.i.d.) manner with no correlation with the adversary's side information.

B. Randomness certification

In this section, we shall certify the randomness that is generated by our protocol. First, we recall two criteria that a randomness generation protocol must satisfy, namely, soundness and completeness.

- (1) *Soundness*

Let Ω be the event in which the protocol is not aborted (i.e., it successfully produces a random output) and let $\Pr[\Omega]$ denotes the probability that this event occurs.

For a given $\varepsilon_{\text{sou}} \in (0, 1)$, a randomness generation protocol is said to be ε_{sou} -sound if the output $\mathbf{K}$ and the adversary's quantum side information E satisfies

$$\Pr[\Omega] \cdot \frac{1}{2} \|\rho_{\mathbf{K}E|\Omega} - \tau_L \otimes \rho_{E|\Omega}\|_1 \leq \varepsilon_{\text{sou}}, \quad (5)$$

where L denotes the length of the output $\mathbf{K}$ and τ_L is the maximally mixed state with dimension 2^L . $\rho_{\mathbf{K}E|\Omega}$ denotes the state of the output of an implementation of the protocol which is not aborted.

- (2) *Completeness*

Let $\varepsilon_{\text{com}} \in (0, 1)$ be given. A randomness generation protocol is said to be ε_{com} -complete if

$$\Pr[\Omega]_{\text{honest}} \geq 1 - \varepsilon_{\text{com}}. \quad (6)$$

Here, the subscript "honest" indicates that the probability of the protocol not being aborted is evaluated assuming that the device behaves honestly (i.e., the device behaves as expected).

The soundness criterion formalizes the intuitive goal that a randomness generation protocol should produce an output that is uniform and independent of the adversary's side information. However, in a self-testing randomness

generation protocol, we must allow the protocol to be aborted when the protocol detects that the output does not have sufficient entropy (e.g., due to some malfunction in the measurement device). Therefore, the soundness criterion demands that either the protocol aborts with high probability or the output is close to being uniform and independent from any side information.

The completeness criterion formalizes the desired property that when the device is behaving honestly, the probability of aborting the protocol should be small. When the device is behaving honestly, the protocol can still be aborted due to statistical fluctuations in the parameter estimation step. In our case, the honest behavior is described by some homodyne measurement, which performs X and P quadrature measurements with known detection efficiency.

Before we present the randomness certification, we shall explicitly state the assumptions that we make to certify the generated randomness. Throughout this work, we assume the following:

- (1) Quantum theory is correct
- (2) The device consists of a trusted and characterized source of quantum states (which we personify as Alice), an uncharacterized measurement device (which we personify as Bob) and a trusted classical processing unit which supplies random inputs, stores measurement outcomes, and performs the necessary postprocessing as specified by the protocol.
- (3) The device is located in a secure location and is sufficiently isolated from the environment during the execution of the protocol. We allow the device to share some correlation (which can be classical or quantum) with the adversary prior to the protocol execution.
- (4) The device is equipped with some trusted random seed that is uncorrelated to the adversary's side information, nor to the devices (i.e., the source and the measurement device) themselves.

The proof for the protocol's completeness is considerably more straightforward. When the device's honest behavior is characterized, the expected score distribution is known. In this case, the protocol will only abort due to statistical fluctuations, and hence to ensure completeness, we only need to set δ to be large enough to allow for these fluctuations.

Let us denote the expected score distribution for the honest behavior as $\mathbf{p}^{\text{hon}} := (1 - \gamma, \gamma \omega^{\text{hon}})$. Here, the first entry refers to the probability of obtaining $C = \perp$, whereas the ω^{hon} gives the conditional score distribution given that the round is a test round. Denote the range of C as $\mathcal{C}$ and we write $\mathcal{C}' = \mathcal{C} \setminus \{\perp\}$ as the range of C for the test rounds. Our strategy to prove the completeness of the protocol is

by deriving an upper bound

$$\Pr[f(c) > \gamma(\omega_c^{\text{hon}} + \delta_c)] \leq \varepsilon_{\text{com},c}, \quad (7)$$

for each $c \in \mathcal{C}'$. Then, the union bound implies that

$$\varepsilon_{\text{com}} \leq \sum_{c \in \mathcal{C}'} \varepsilon_{\text{com},c}. \quad (8)$$

In this work, to obtain the upper bound $\varepsilon_{\text{com},c}$ in Eq. (7), we use the following strategy. For each $c \in \mathcal{C}'$, we consider a sequence of random variables that indicate whether for a given round i , Alice and Bob obtain the score $C_i = c$. For each i , this random variable is a Bernoulli random variable, and since the honest behavior between rounds will be independent, the number of rounds with score c follows a binomial distribution, with n trials and probability $\gamma \omega_c^{\text{hon}}$. We can then use the bounds on binomial distribution presented in Ref. [62] to obtain the upper bounds

$$\varepsilon_{\text{com},c} = 1 - F(n, \gamma \omega_c^{\text{hon}}, \lfloor n\gamma(\omega_c^{\text{hon}} + \delta_c) \rfloor), \quad (9)$$

where

$$D(q, p) := q \ln \left(\frac{q}{p} \right) + (1 - q) \ln \left(\frac{1 - q}{1 - p} \right), \quad (10)$$

$$\Phi(a) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a dx \exp \left(-\frac{x^2}{2} \right), \quad (11)$$

$$F(n, p, k) := \Phi \left(\text{sign} \left(\frac{k}{n} - p \right) \sqrt{2nD \left(\frac{k}{n}, p \right)} \right). \quad (12)$$

The union bound (8) concludes the completeness proof.

For soundness, assuming that the adversary's side information is limited by quantum theory, we can invoke leftover hashing against quantum side information [63]. The leftover hash lemma states that we can use a two-universal family of hash functions as a quantum-proof strong extractor **Ext**. If $\mathbf{R}$ is the input of the extractor and $\mathbf{S}$ is a uniformly chosen random seed, and we let $\mathbf{Z} = \text{Ext}(\mathbf{R}, \mathbf{S})$, then the output of the randomness extraction step satisfies the following equation:

$$\begin{aligned} & \frac{1}{2} \|\rho_{\mathbf{ZS}E} - \tau_\ell \otimes \tau_s \otimes \rho_E\|_1 \\ & \leq 2\epsilon_s + 2^{-\frac{1}{2}(H_{\min}^{\epsilon_s}(\mathbf{R}|E) - \ell + 2)}, \end{aligned} \quad (13)$$

for any $\epsilon_s \in (0, \sqrt{\text{Tr}[\rho]})$. Here, ℓ is the length of the output $\mathbf{Z}$ while s is the length of the seed $\mathbf{S}$ that we require for randomness extraction. Thus, the concatenated length $L = \ell + s$.

If we set the right-hand side of the inequality (13) to ε_{sou} , then it would be sufficient to prove a lower bound on the

conditional smooth min entropy $H_{\min}^{\epsilon_s}(\mathbf{R}|E)_{\rho_{RE}}$ to establish the soundness of the protocol. However, in practice, it is not easy to establish a lower bound on the conditional smooth min-entropy directly. Instead, we consider two cases (that are non-exhaustive):

- (a) $\Pr[\Omega] < \epsilon_{EA}$ for some fixed $\epsilon_{EA} \in (0, 1)$
- (b) $\Pr[\Omega] \geq \epsilon_{EA}$ and $H_{\min}^{\epsilon_s}(\mathbf{R}|E)_{\rho_{RE \wedge \Omega}} \geq k$ for some constant k that we will determine later.

Here, $\rho_{RE \wedge \Omega} = \Pr[\Omega] \cdot \rho_{RE|\Omega}$ is a subnormalized state associated with the case where the protocol is not aborted.

For case (a), we can use $\epsilon_{\text{sou}} = \epsilon_{EA}$ as an upper bound for the soundness criterion (5). For case (b), we can use k as a lower bound the conditional smooth min-entropy and invoke the leftover hash lemma. If we set the exponential term of Eq. (13) to be equal to ϵ_{ext} , then we can set $\epsilon_{\text{sou}} = 2\epsilon_s + \epsilon_{\text{ext}}$ and the maximum output length that we can have is given by

$$\ell = \left\lfloor k - 2 \log_2 \left(\frac{1}{\epsilon_{\text{ext}}} \right) + 2 \right\rfloor. \quad (14)$$

Combining the two cases, we conclude that the soundness parameter can be set as

$$\epsilon_{\text{sou}} = \max\{\epsilon_{EA}, 2\epsilon_s + \epsilon_{\text{ext}}\}. \quad (15)$$

Now, our task is to find the constant k , which is a lower bound on the conditional smooth min-entropy in case (b). Recall that in our protocol, the input $\mathbf{R}$ to the randomness extractor $\mathbf{Ext}$ is given by the concatenation of the measurement outcomes $\mathbf{B}$ and the inputs $(\mathbf{T}, \mathbf{X}, \mathbf{Y})$. The conditional smooth min-entropy can be lower bounded using the chain rule for smooth entropies [64]

$$\begin{aligned} H_{\min}^{\epsilon_s}(\mathbf{R}|E)_{\rho_{\wedge \Omega}} &\geq H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TX}\mathbf{Y}, E)_{\rho_{\wedge \Omega}} + H_{\min}^{\epsilon_2}(\mathbf{TX}\mathbf{Y}|E)_{\rho_{\wedge \Omega}} \\ &\quad - \log_2 \left(\frac{2}{\epsilon_s - \epsilon_2 - 2\epsilon_1} \right). \end{aligned} \quad (16)$$

Lower bounding the second term is easy since we have assumed that the inputs are chosen from a trusted random seed, which is also independent of the adversary's side information E . Furthermore, the inputs are identical and independently distributed (IID), which allows us to invoke the quantum asymptotic equipartition property (QAP) [65]. The QAP [66] states that

$$H_{\min}^{\epsilon_2}(\mathbf{TX}\mathbf{Y}|E)_{\rho} \geq nH(T_i, X_i, Y_i|E) - \xi\sqrt{n}, \quad (17)$$

where

$$\xi = 2 \log_2(1 + 4m)\sqrt{1 - 2 \log_2 \epsilon_2}. \quad (18)$$

From our assumption that the inputs are independent from E , we have

$$\begin{aligned} H(T_i, X_i, Y_i|E) &= H(T_i, X_i, Y_i) \\ &= H(T_i) + H(X_i, Y_i|T_i) \\ &= H(T_i) + \gamma H(X_i, Y_i|T_i = 1) \\ &= h_2(\gamma) + 2\gamma, \end{aligned} \quad (19)$$

where in the first line, we use the independence from E , in the second line, we use the chain rule for Shannon entropy, in the third line, we decompose the conditional entropy term and drop the $H(X_i, Y_i|T_i = 0)$ term since the inputs are fixed in the generation rounds. In the last line, we compute the value of each term, where h_2 denotes the binary entropy function.

While we have obtained the conditional smooth min-entropy $H_{\min}^{\epsilon_2}(\mathbf{TX}\mathbf{Y}|E)_{\rho}$ evaluated on the unconditioned state $\rho_{\mathbf{TX}\mathbf{Y}E}$, we can convert this to the state $\rho_{\mathbf{TX}\mathbf{Y}E \wedge \Omega}$ using Lemma 10 of Ref. [67] as long as $\epsilon_2 \in (0, \sqrt{\Pr[\Omega]})$. This can be done by choosing $\epsilon_2 < \epsilon_{EA} < \sqrt{\epsilon_{EA}} \leq \sqrt{\Pr[\Omega]}$ since we are in case (b). Thus, we have established the bound

$$H_{\min}^{\epsilon_2}(\mathbf{TX}\mathbf{Y}|E)_{\rho_{\wedge \Omega}} \geq n[h_2(\gamma) + 2\gamma] - \xi\sqrt{n}. \quad (20)$$

For the first term of (16), we can use the entropy accumulation theorem (EAT) [54,55] to bound $H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TX}\mathbf{Y}, E)_{\rho_{\wedge \Omega}}$ in case (b). EAT relates the conditional smooth min-entropy for the entire protocol (i.e., accumulated over n rounds) to a single-round quantity called the min-trade-off function f_v (for the precise definition, we refer the readers to Refs. [54,55,68]). This makes the analysis tractable as the min-trade-off function can be constructed using semidefinite programming (SDP) method [18,68,69].

There is some condition to the application of EAT, namely that the inputs and outputs satisfy the so-called Markov chain condition [54,55]. Based on the same argument as the one presented in our previous experiment [18], this is satisfied in our experiment as the inputs are chosen from an independent and trusted random seed and the device is sufficiently isolated such that the adversary's quantum side information is not updated in the protocol.

To construct the min-trade-off function, let $P_{\text{guess}}(\omega)$ denotes the single-round guessing probability subject to the score constraints ω . Formally, it is defined as the solution to the following optimization:

$$\begin{aligned} \max_{\{M_{b|y}\}, \{E_e\}} &\sum_{b=0}^1 \langle \varphi_0 | M_{b|1} \otimes E_b | \varphi_0 \rangle \\ \text{s.t.} &\sum_{b,x,y} \kappa_{b,x,y}^{(c)} \langle \varphi_x | M_{b|y} \otimes \mathbb{I} | \varphi_x \rangle = \omega_c, \end{aligned} \quad (21)$$

where $|\varphi_x\rangle$ is the state shared by Bob and the adversary (which includes any pre-shared entanglement between

them). Since we assume the correctness of quantum theory, we can describe measurements using positive operator-valued measure (POVM). Here, $\{M_{b|y}\}_{b,y}$ is the POVM elements describing Bob's untrusted measurement and $\{E_e\}_e$ is the POVM elements describing the adversary's measurement. $\kappa_{b,x,y}^{(c)}$ are the coefficients that gives the appropriate weight to the game's scores.

Let α_v and λ_v be constants such that

$$P_{\text{guess}}(\omega) \leq \alpha_v + \lambda_v \cdot \omega \quad (22)$$

for any score distribution ω that is compatible with quantum theory. The constants α_v and λ_v can be found by taking the dual optimization (21) for $P_{\text{guess}}(v)$ (which motivates the subscripts on α_v and λ_v).

Following the derivation of Theorem 1 presented in Ref. [18], EAT implies that the conditional smooth min-entropy $H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TXY}, E)_{\rho_{|\Omega}}$ is given by

$$H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TXY}, E)_{\rho_{|\Omega}} \geq nh - n[\beta V + \beta^2 K] - \frac{1}{\beta}[1 - 2 \log_2(\epsilon_{\text{EA}} \epsilon_1)], \quad (23)$$

where

$$h = 2(1 - \gamma)(1 - \alpha_v - \lambda_v \cdot \tilde{\omega}), \quad (24)$$

$$V = \frac{\ln 2}{2} \left(\log_2(4m + 1) + \sqrt{2 + \frac{4(1 - \gamma)^2(\lambda_{\max} - \lambda_{\min})^2}{\gamma}} \right)^2, \quad (25)$$

$$K = \frac{2^{\beta(\log_2 2m + 2(1 - \gamma)(\lambda_{\max} - \lambda_{\min}))}}{6 \ln 2(1 - \beta)^3} \times \ln^3(2^{\log_2 2m + 2(1 - \gamma)(\lambda_{\max} - \lambda_{\min})} + e^2), \quad (26)$$

where $\lambda_{\max} = \max_c \lambda_v$ and $\lambda_{\min} = \min_c \lambda_v$. Letting $c_{\min} = \text{argmin}_c \lambda_v$, we define $\tilde{\omega}$ as

$$\tilde{\omega}_c = \begin{cases} \omega_c + \delta_c & \text{if } c \neq c_{\min}, \perp, \\ \omega_c - \sum_{c \neq c_{\min}, \perp} \delta_c & \text{if } c = c_{\min}. \end{cases} \quad (27)$$

Thus, we have a lower bound on the conditional smooth min-entropy $H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TXY}, E)_{\rho_{|\Omega}}$ evaluated on the state $\rho_{\mathbf{B}|\mathbf{TXY}E|\Omega}$. However, we are interested in the conditional smooth min-entropy evaluated on the *subnormalized* version of the state $\rho_{\mathbf{B}|\mathbf{TXY}E\wedge\Omega}$. We shall show that we indeed have $H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TXY}, E)_{\rho_{\wedge\Omega}} \geq H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TXY}, E)_{\rho_{|\Omega}}$.

First, recall that given the state ρ_{AB} , the conditional min-entropy can be defined as [70]

$$H_{\min}(A|B)_{\rho} = \max_{\sigma_B} \sup\{\zeta : \rho_{AB} \preceq 2^{-\zeta} \mathbb{I} \otimes \sigma_B\}. \quad (28)$$

Meanwhile, conditional smooth min-entropy is defined as [70]

$$H_{\min}^{\epsilon}(A|B)_{\rho} = \max_{\hat{\rho}: \mathcal{P}(\rho, \hat{\rho}) \leq \epsilon} H_{\min}(A|B)_{\hat{\rho}}, \quad (29)$$

where $\mathcal{P}(\rho, \hat{\rho})$ denotes the purified distance between the state ρ and $\hat{\rho}$, considering the above two definitions, if $H_{\min}^{\epsilon}(A|B)_{\rho} = h_*$, then there exist state $\hat{\rho}_{AB}$ and state σ_B such that

$$\hat{\rho}_{AB} \preceq 2^{-h_*} \mathbb{I}_A \otimes \sigma_B \quad (30)$$

and

$$\mathcal{P}(\rho_{AB}, \hat{\rho}_{AB}) \leq \epsilon. \quad (31)$$

Now, let $0 < p \leq 1$. Let $\rho' = p\rho$ and $\hat{\rho}' = p\hat{\rho}$. Clearly, multiplying each state by p is a completely positive, trace non-increasing (CPTNI) map. Due to monotonicity of purified distance under CPTNI maps, we have

$$\mathcal{P}(\rho'_{AB}, \hat{\rho}'_{AB}) \leq \mathcal{P}(\rho_{AB}, \hat{\rho}_{AB}) \leq \epsilon. \quad (32)$$

Furthermore, we have

$$\hat{\rho}'_{AB} \preceq \hat{\rho}_{AB} \preceq 2^{-h_*} \mathbb{I}_A \otimes \sigma_B. \quad (33)$$

This implies that $H_{\min}^{\epsilon}(A|B)_{\rho'} \geq h_*$ as well. In other words, multiplying a state by a number $0 < p \leq 1$ does not reduce the conditional smooth min-entropy. We apply this property to our case where the initial state is $\rho_{\mathbf{B}|\mathbf{TXY}E|\Omega}$ and the final state is $\rho_{\mathbf{B}|\mathbf{TXY}E\wedge\Omega} = \text{Pr}[\Omega] \cdot \rho_{\mathbf{B}|\mathbf{TXY}E|\Omega}$.

Therefore, we have lower bounds for both $H_{\min}^{\epsilon_1}(\mathbf{B}|\mathbf{TXY}, E)_{\rho_{\wedge\Omega}}$ and $H_{\min}^{\epsilon_2}(\mathbf{TXY}|E)_{\rho_{\wedge\Omega}}$. We can then use the chain rule (16) to derive the lower bound k for case (b)

$$\begin{aligned} k &= n[h + h_2(\gamma) + 2\gamma] \\ &\quad - n[\beta V + \beta^2 K] - \frac{1}{\beta}[1 - 2 \log_2(\epsilon_{\text{EA}} \epsilon_1)] \\ &\quad - \xi \sqrt{n} - \log_2 \left(\frac{2}{\epsilon_s - \epsilon_2 - 2\epsilon_1} \right). \end{aligned} \quad (34)$$

This concludes the soundness proof of the protocol.

Now, to calculate the expected randomness expansion rate, we have to consider the amount of randomness that we need to generate the inputs. There are some efficient algorithms (such as the interval algorithm [71]) that convert a uniform bit strings into biased bit strings of shorter

length, which we need for our protocol. For the interval algorithm, the required input bit string has a length of

$$\ell_{\text{in}} = nH(T_i, X_i, Y_i) + 3 = n[h_2(\gamma) + 2\gamma] + 3. \quad (35)$$

We compare this to the output length

$$\begin{aligned} \ell = & \left\lfloor n[h + h_2(\gamma) + 2\gamma] \right. \\ & - n(\beta V + \beta^2 K) - \frac{1}{\beta} [1 - 2 \log_2(\epsilon_{\text{EA}} \epsilon_1)] - \xi \sqrt{n} \\ & \left. - \log_2 \left(\frac{2}{\epsilon_s - \epsilon_2 - 2\epsilon_1} \right) - 2 \log_2 \left(\frac{1}{\epsilon_{\text{ext}}} \right) + 2 \right\rfloor. \end{aligned} \quad (36)$$

$$(37)$$

This gives the expected randomness expansion rate

$$\begin{aligned} \mathbf{r}_{\text{net}} = & h - \beta(V + \beta K) - \frac{1 - 2 \log_2(\epsilon_{\text{EA}} \epsilon_1)}{n\beta} - \frac{\xi}{\sqrt{n}} \\ & - \frac{\log_2 \left(\frac{2}{\epsilon_s - \epsilon_2 - 2\epsilon_1} \right) + 2 \log_2 \left(\frac{1}{\epsilon_{\text{ext}}} \right) - 1}{n}. \end{aligned} \quad (38)$$

To obtain the scaling that we claimed in Eq. (1), we can choose $\beta \sim \mathcal{O}(1/\sqrt{n})$. However, in our plots, we optimize β in Eq. (38) to maximize the expected randomness expansion rate.

The security of our protocol is guaranteed by the quantum-proof strong extractor, such that either the protocol is aborted with high probability (e.g., in the case where the device does not behave as expected), or the output string is close to an ideal random string, which is uniformly distributed and independent from any quantum side information and initial random seeds.

C. Experimental setup and data collection procedure

The photonic chip is wire-bond packaged on a customized PCB that hosts all the electrical input-output ports. The direct current (DC) ports are controlled by an off-chip multi-channel power supply (Qontrol Q8iv). The radio frequency (RF) ports are connected to arbitrary waveform generators (Zurich Instrument HDAWG4). The homodyne detector amplifier is customized based on the Texas Instruments OPA847 operational amplifier chip. The photonic chip is mounted close to the amplifier chip to minimize any input parasitics. The output signal from the amplifier is collected using an oscilloscope (LeCroy WavePro HD) and analyzed on a computer.

The optical input is an O-band tunable laser (EXFO T100S) with a maximum output power of 14.5 mW. The laser passes a polarization controller and is coupled to the photonic chip with the assistance of a precision alignment stage. The output light splits into two paths. One

part is measured using an off-chip power meter (Thorlab PM100D), and the other connects to a photodetector for waveform analysis.

The chip contains multiple heat sources, including the thermal optical phase shifter and MZM. The thermal optical phase shifters are used mainly to address the fabrication errors on the chip, and the MZM has a high-speed driving signal that is higher than the thermal response time. Thus, both can be treated as constant heat sources after a few minutes of warm-up. Heat crosstalk on the chip is measured to be negligible. Another crosstalk we observed is the electromagnetic crosstalk, from both the environment and chip components. We designed an electromagnetic shielding using a thick layer (1–10 μm) of aluminum or copper covering both the top and bottom surfaces of the chip. The crosstalk is negligible with such a shielding layer applied.

Considering the large number of rounds required, the system settings, including all the bias control and fiber-to-chip alignment, are monitored and calibrated constantly (every 100 s) to avoid any environmental drift (mainly the temperature drift and fiber-chip alignment drift). The homodyne tunable BS is fine-tuned first with the signal path set to a high attenuation state and the LO path with an arbitrary modulation. The modulation on LO is minimized when the CMRR is at the optimum state. Then, the LO phase modulator P bias is fine-tuned to ensure the intensity on the X and P quadratures is the same. The IQ modulator biases are calibrated next to ensure accurate QPSK states. Finally, the attenuators are used to set the state intensity. The resulting intensity variation on the LO path is within about $\pm 2.2\%$ and the phase error is within $\pm 0.6\%$.

The whole calibration process is repeated every 100 s (approximately), which is the time for 50 rounds of oscilloscope data collection. The frequent calibration process ensures the high requirement of state intensity. The transfer and calibration time reduced the actual average sample rate by 100 times, to about 100 kSample/s.

Active temperature control using a Peltier cooler could significantly reduce the requirement for frequent recalibration. Advanced optoelectronic packaging could also allow for the use of fiber array inputs that can provide more optical feedback control ports, potentially reducing the calibration time. The data transfer bottleneck can be improved by using high-speed FPGA-based data collection cards.

In the experiment, the input generation step uses a pseudorandom input with a length of 100. Both the X and P quadrature for each state are measured to assist the system characterization. The unused measurements are discarded before the postprocessing process. The data are collected and digitally processed in blocks. The oscilloscope collects data with a memory size of 50 MSamples. After each calibration step, the oscilloscope memory is recorded 50 times to form a block, which corresponds to 2.5×10^7 number of rounds. All the data within the 2.5×10^7 rounds block are used to calculate the input state amplitude, whose

accuracy is crucial in operating the protocol. We set a range of ± 0.0022 for the acceptable amplitude. The whole data block is discarded if the amplitude exceeds the acceptable range.

ACKNOWLEDGMENTS

The authors acknowledge funding support from the Zhejiang Provincial Natural Science Foundation of China (No. Z25F050014), National Natural Science Foundation of China (NSFC) (No. 62505275), Zhejiang University Education Foundation Qizhen Scholar Foundation, the National Research Foundation of Singapore (NRF) Fellowship grant (No. NRFF11-2019-0001), NRF Quantum Engineering Programme 1.0 grant (Nos. QEP-P2 and QEP-P3), MOE Tier 1 (No. A-8001168-00-00), and NRF-A*STAR Quantum Engineering Programme (No. NRF2021-QEP2-04-P01).

C.L. and G.Z. jointly conceived the idea. C.W. and G.Z. designed the photonic chips and experiments. C.L., I.W.P., and K.T.G. designed the randomness expansion protocol with the randomness certification proof. C.Y. and G.X. supported the photonic chip packaging. C.W., H.J.N., S.Q.N., and Z.G. conducted the experiments. All authors contributed to the discussion of experimental results. C.L. supervised and coordinated all the work. G.Z. and I.W.P. wrote the manuscript with contributions from all co-authors.

The authors declare no competing financial interests.

This paper was prepared for informational purposes with contributions from the Global Technology Applied Research center of JPMorganChase. This paper is not a product of the Research Department of JPMorganChase. or its affiliates. Neither JPMorganChase. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this position paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

DATA AVAILABILITY

All of the data that support the findings of this study are available in the main text. Source data are available from the corresponding author on request.

[1] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, N. Hecker, J.

Dray, S. Vo, and L. Bassham, A statistical test suite for random and pseudorandom number generators for cryptographic applications, Technical Report Special Publication 800-22, Revision 1a (National Institute of Standards and Technology, Gaithersburg, MD, 2010).

[2] M. Stipčević and Ç. K. Koç, in *Open Problems in Mathematics and Computational Science*, edited by Ç. K. Koç (Springer International Publishing, Cham, 2014), p. 275.

[3] M. Herrero-Collantes and J. C. Garcia-Escartin, Quantum random number generators, *Rev. Mod. Phys.* **89**, 015004 (2017).

[4] X. Ma, X. Yuan, Z. Cao, B. Qi, and Z. Zhang, Quantum random number generation, *npj Quantum Inf.* **2**, 1 (2016).

[5] Y.-K. Liu and D. Moody, Post-quantum cryptography and the quantum future of cybersecurity, *Phys. Rev. Appl.* **21**, 040501 (2024).

[6] D. Mayers and A. Yao, Self testing quantum apparatus, *Quantum Inf. Comput.* **4**, 273 (2004).

[7] I. Šupić and J. Bowles, Self-testing of quantum systems: A review, *Quantum* **4**, 337 (2020).

[8] A. Acín and L. Masanes, Certified randomness in quantum physics, *Nature* **540**, 213 (2016).

[9] S. Pironio, A. Acín, S. Massar, A. B. de La Giroday, D. N. Matsukevich, P. Maunz, S. Olmschenk, D. Hayes, L. Luo, T. A. Manning *et al.*, Random numbers certified by Bell's theorem, *Nature* **464**, 1021 (2010).

[10] W.-Z. Liu, M.-H. Li, S. Ragy, S.-R. Zhao, B. Bai, Y. Liu, P. J. Brown, J. Zhang, R. Colbeck, J. Fan *et al.*, Device-independent randomness expansion against quantum side information, *Nat. Phys.* **17**, 448 (2021).

[11] L. K. Shalm, Y. Zhang, J. C. Bienfang, C. Schlager, M. J. Stevens, M. D. Mazurek, C. Abellán, W. Amaya, M. W. Mitchell, M. A. Alhejji *et al.*, Device-independent randomness expansion with entangled photons, *Nat. Phys.* **17**, 452 (2021).

[12] D.-L. Deng, C. Zu, X.-Y. Chang, P.-Y. Hou, H.-X. Yang, Y.-X. Wang, and L.-M. Duan, Exploring quantum contextuality to generate true random numbers, *ArXiv:1301.5364*.

[13] V. C. Vivoli, P. Sekatski, J.-D. Bancal, C. C. W. Lim, A. Martin, R. T. Thew, H. Zbinden, N. Gisin, and N. Sangouard, Comparing different approaches for generating random numbers device-independently using a photon pair source, *New J. Phys.* **17**, 023023 (2015).

[14] D. G. Marangon, G. Vallone, and P. Villoresi, Source-device-independent ultrafast quantum random number generation, *Phys. Rev. Lett.* **118**, 060503 (2017).

[15] Z. Cao, H. Zhou, X. Yuan, and X. Ma, Source-independent quantum random number generation, *Phys. Rev. X* **6**, 011020 (2016).

[16] M. Avesani, D. G. Marangon, G. Vallone, and P. Villoresi, Source-device-independent heterodyne-based quantum random number generator at 17 Gbps, *Nat. Commun.* **9**, 5365 (2018).

[17] J.-N. Zhang, R. Yang, X. Li, C.-W. Sun, Y.-C. Liu, Y. Wei, J.-C. Duan, Z. Xie, Y.-X. Gong, and S.-N. Zhu, Realization of a source-device-independent quantum random number generator secured by nonlocal dispersion cancellation, *Adv. Photonics* **5**, 036003 (2023).

[18] C. Wang, I. W. Primaatmaja, H. J. Ng, J. Y. Haw, R. Ho, J. Zhang, G. Zhang, and C. Lim, Provably-secure quantum

- randomness expansion with uncharacterised homodyne detection, *Nat. Commun.* **14**, 316 (2023).
- [19] Y.-Q. Nie, J.-Y. Guan, H. Zhou, Q. Zhang, X. Ma, J. Zhang, and J.-W. Pan, Experimental measurement-device-independent quantum random-number generation, *Phys. Rev. A* **94**, 060301(R) (2016).
- [20] Z. Cao, H. Zhou, and X. Ma, Loss-tolerant measurement-device-independent quantum random number generation, *New J. Phys.* **17**, 125011 (2015).
- [21] J. B. Brask, A. Martin, W. Esposito, R. Houlmann, J. Bowles, H. Zbinden, and N. Brunner, Megahertz-rate semi-device-independent quantum random number generators based on unambiguous state discrimination, *Phys. Rev. Appl.* **7**, 054018 (2017).
- [22] T. Lunghi, J. B. Brask, C. C. W. Lim, Q. Lavigne, J. Bowles, A. Martin, H. Zbinden, and N. Brunner, Self-testing quantum random number generator, *Phys. Rev. Lett.* **114**, 150501 (2015).
- [23] Y. Zhang, H.-P. Lo, A. Mink, T. Ikuta, T. Honjo, H. Takesue, and W. J. Munro, A simple low-latency real-time certifiable quantum random number generator, *Nat. Commun.* **12**, 1056 (2021).
- [24] D. Rusca, T. Van Himbeek, A. Martin, J. B. Brask, W. Shi, S. Pironio, N. Brunner, and H. Zbinden, Self-testing quantum random-number generator based on an energy bound, *Phys. Rev. A* **100**, 062338 (2019).
- [25] M. Avesani, H. Tebyanian, P. Villoresi, and G. Vallone, Semi-device-independent heterodyne-based quantum random-number generator, *Phys. Rev. Appl.* **15**, 034034 (2021).
- [26] C. Bruynsteen, T. Gehring, C. Lupo, J. Bauwelinck, and X. Yin, 100-Gbit/s integrated quantum random number generator based on vacuum fluctuations, *PRX Quantum* **4**, 010330 (2023).
- [27] B. Bai, J. Huang, G.-R. Qiao, Y.-Q. Nie, W. Tang, T. Chu, J. Zhang, and J.-W. Pan, 18.8 Gbps real-time quantum random number generator with a photonic integrated chip, *Appl. Phys. Lett.* **118**, 264001 (2021).
- [28] T. Roger, T. Paraiso, I. D. Marco, D. G. Marangon, Z. Yuan, and A. J. Shields, Real-time interferometric quantum random number generation on chip, *J. Opt. Soc. Am. B* **36**, B137 (2019).
- [29] A. Stanco, D. G. Marangon, G. Vallone, S. Burri, E. Charbon, and P. Villoresi, Efficient random number generation techniques for CMOS single-photon avalanche diode array exploiting fast time tagging units, *Phys. Rev. Res.* **2**, 023287 (2020).
- [30] T. Bertapelle, M. Avesani, A. Santamato, A. Montanaro, M. Chiesa, D. Rotta, M. Artiglia, V. Sorianello, F. Testa, G. De Angelis *et al.*, High-speed source-device-independent quantum random number generator on a chip, *Opt. Quantum* **3**, 111 (2025).
- [31] P. Kincaid, L. De Marinis, A. Montanaro, A. Santamato, N. Andriolli, and G. Contestabile, in *2023 International Conference on Photonics in Switching and Computing (PSC)* (IEEE, Mantova, Italy, 2023), p. 1.
- [32] Y. Du, X. Hua, Z. Zhao, X. Sun, Z. Zhang, X. Xiao, and K. Wei, Source-independent quantum random number generators with integrated silicon photonics, *Commun. Phys.* **8**, 9 (2025).
- [33] N. Leone, D. Rusca, S. Azzini, G. Fontana, F. Acerbi, A. Gola, A. Tontini, N. Massari, H. Zbinden, and L. Pavesi, An optical chip for self-testing quantum random number generation, *APL Photonics* **5**, 101301 (2020).
- [34] B. Haylock, D. Peace, F. Lenzini, C. Weedbrook, and M. Lobino, Multiplexed quantum random number generation, *Quantum* **3**, 141 (2019).
- [35] C. Abellan, W. Amaya, D. Domenech, P. Muñoz, J. Capmany, S. Longhi, M. W. Mitchell, and V. Pruneri, Quantum entropy source on an InP photonic integrated circuit for random number generation, *Optica* **3**, 989 (2016).
- [36] A. Prokhorov, V. Kovalyuk, P. An, A. Golikov, R. Shakhovoy, V. Sharoglazova, A. Udaltsov, Y. Kurochkin, and G. Goltsman, Silicon nitride Mach-Zehnder interferometer for on-chip quantum random number generation, *J. Phys.: Conf. Ser.* **1695**, 012118 (2020).
- [37] F. Raffaelli, P. Sibson, J. E. Kennard, D. H. Mahler, M. G. Thompson, and J. C. Matthews, Generation of random numbers by measuring phase fluctuations from a laser diode with a silicon-on-insulator chip, *Opt. Express* **26**, 19730 (2018).
- [38] J. Argillander, A. Alarcón, C. Bao, C. Kuang, G. Lima, F. Gao, and G. B. Xavier, Quantum random number generation based on a perovskite light emitting diode, *Commun. Phys.* **6**, 157 (2023).
- [39] A. H. Atabaki, S. Moazeni, F. Pavanello, H. Gevorgyan, J. Notaros, L. Alloatti, M. T. Wade, C. Sun, S. A. Kruger, H. Meng *et al.*, Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip, *Nature* **556**, 349 (2018).
- [40] J. Wang, F. Sciarrino, A. Laing, and M. G. Thompson, Integrated photonic quantum technologies, *Nat. Photonics* **14**, 273 (2020).
- [41] W. Luo, L. Cao, Y. Shi, L. Wan, H. Zhang, S. Li, G. Chen, Y. Li, S. Li, Y. Wang *et al.*, Recent progress in quantum photonic chips for quantum communication and internet, *Light: Sci. Appl.* **12**, 175 (2023).
- [42] M. Zhang, L. Feng, M. Li, Y. Chen, L. Zhang, D. He, G. Guo, G. Guo, X. Ren, and D. Dai, Supercompact photonic quantum logic gate on a silicon chip, *Phys. Rev. Lett.* **126**, 130501 (2021).
- [43] C. Vigliar, S. Paesani, Y. Ding, J. C. Adcock, J. Wang, S. Morley-Short, D. Bacco, L. K. Oxenløwe, M. G. Thompson, J. G. Rarity *et al.*, Error-protected qubits in a silicon photonic chip, *Nat. Phys.* **17**, 1137 (2021).
- [44] X. Qiang, Y. Wang, S. Xue, R. Ge, L. Chen, Y. Liu, A. Huang, X. Fu, P. Xu, T. Yi *et al.*, Implementing graph-theoretic quantum algorithms on a silicon photonic quantum walk processor, *Sci. Adv.* **7**, eabb8375 (2021).
- [45] K. Wei, W. Li, H. Tan, Y. Li, H. Min, W.-J. Zhang, H. Li, L. You, Z. Wang, X. Jiang *et al.*, High-speed measurement-device-independent quantum key distribution with integrated silicon photonics, *Phys. Rev. X* **10**, 031030 (2020).
- [46] P. Sibson, J. E. Kennard, S. Stanisic, C. Erven, J. L. O'Brien, and M. G. Thompson, Integrated silicon photonics for high-speed quantum key distribution, *Optica* **4**, 172 (2017).
- [47] G. Zhang, J. Y. Haw, H. Cai, F. Xu, S. Assad, J. F. Fitzsimons, X. Zhou, Y. Zhang, S. Yu, J. Wu *et al.*, An

- integrated silicon photonic chip platform for continuous-variable quantum key distribution, *Nat. Photonics* **13**, 839 (2019).
- [48] D. Bunandar, A. Lentine, C. Lee, H. Cai, C. M. Long, N. Boynton, N. Martinez, C. DeRose, C. Chen, M. Grein *et al.*, Metropolitan quantum key distribution with silicon photonics, *Phys. Rev. X* **8**, 021009 (2018).
- [49] C. Ma, W. D. Sacher, Z. Tang, J. C. Mikkelsen, Y. Yang, F. Xu, T. Thiessen, H.-K. Lo, and J. K. Poon, Silicon photonic transmitter for polarization-encoded quantum key distribution, *Optica* **3**, 1274 (2016).
- [50] J. Wang, M. Al-Qadi, W.-J. Jiang, K. Li, Y.-W. Chen, K. Kuzmin, C. Ho, Y. Y. Zhao, H. Yamada, J. Ackert *et al.*, 120-GBaud 16-QAM silicon photonics IQ modulator for data center interconnection, *Opt. Express* **31**, 25515 (2023).
- [51] J. S. Bell, On the Einstein-Podolsky-Rosen paradox, *Phys. Phys. Fiz.* **1**, 195 (1964).
- [52] N. Brunner, D. Cavalcanti, S. Pironio, V. Scarani, and S. Wehner, Bell nonlocality, *Rev. Mod. Phys.* **86**, 419 (2014).
- [53] I. W. Primaatmaja, K. T. Goh, E. Y.-Z. Tan, J. T.-F. Khoo, S. Ghorai, and C. C.-W. Lim, Security of device-independent quantum key distribution protocols: A review, *Quantum* **7**, 932 (2023).
- [54] F. Dupuis and O. Fawzi, Entropy accumulation with improved second-order term, *IEEE Trans. Inf. Theory* **65**, 7596 (2019).
- [55] F. Dupuis, O. Fawzi, and R. Renner, Entropy accumulation, *Commun. Math. Phys.* **379**, 867 (2020).
- [56] S. Q. Ng, G. Zhang, C. Lim, and C. Wang, A chip-integrated homodyne detection system with enhanced bandwidth performance for quantum applications, *Quantum Sci. Technol.* **9**, 045010 (2024).
- [57] S. Q. Ng, F. Kanitschar, G. Zhang, and C. Wang, Gigabit-rate quantum key distribution on integrated photonic chips, *Optica* **13** (2026).
- [58] A. Marinins, S. Hänsch, H. Sar, F. Chancerel, N. Golshani, H.-L. Wang, A. Tsiara, D. Coenen, P. Verheyen, G. Capuz *et al.*, Wafer-scale hybrid integration of InP DFB lasers on Si photonics by flip-chip bonding with sub-300 nm alignment precision, *IEEE J. Sel. Top. Quantum Electron* **29**, 1 (2022).
- [59] N. Li, G. Chen, D. K. Ng, L. W. Lim, J. Xue, C. P. Ho, Y. H. Fu, and L. Y. Lee, Integrated lasers on silicon at communication wavelength: A progress review, *Adv. Opt. Mater.* **10**, 2201008 (2022).
- [60] J. Y. S. Tan, S. X. Wu, S. Yanikgonul, C. Li, and P. G.-Q. Lo, Foundry's perspective on laser and SOA module integration with silicon photonics, *J. Lightwave Technol.* **42**, 1062 (2023).
- [61] H. J. Ng, W. Y. Kon, I. W. Primaatmaja, C. Wang, and C. Lim, Sampled subblock hashing for large-input randomness extraction, *Phys. Rev. Appl.* **21**, 034005 (2024).
- [62] A. M. Zubkov and A. A. Serov, A complete proof of universal inequalities for the distribution function of the binomial law, *Theory Probab. Appl.* **57**, 539 (2013).
- [63] M. Tomamichel, C. Schaffner, A. Smith, and R. Renner, Leftover hashing against quantum side information, *IEEE Trans. Inf. Theory* **57**, 5524 (2011).
- [64] A. Vitanov, F. Dupuis, M. Tomamichel, and R. Renner, Chain rules for smooth min- and max-entropies, *IEEE Trans. Inf. Theory* **59**, 2603 (2013).
- [65] M. Tomamichel, R. Colbeck, and R. Renner, A fully quantum asymptotic equipartition property, *IEEE Trans. Inf. Theory* **55**, 5840 (2009).
- [66] We use the version of QAEP given in Corollary 4.10 of Ref. [55].
- [67] M. Tomamichel and A. Leverrier, A largely self-contained and complete security proof for quantum key distribution, *Quantum* **1**, 14 (2017).
- [68] P. J. Brown, S. Ragy, and R. Colbeck, A framework for quantum-secure device-independent randomness expansion, *IEEE Trans. Inf. Theory* **66**, 2964 (2019).
- [69] Y. Wang, I. W. Primaatmaja, E. Lavie, A. Varvitsiotis, and C. C. W. Lim, Characterising the correlations of prepare-and-measure quantum networks, *npj Quantum Inf.* **5**, 17 (2019).
- [70] M. Tomamichel, R. Colbeck, and R. Renner, Duality between smooth min- and max-entropies, *IEEE Trans. Inf. Theory* **56**, 4674 (2010).
- [71] Te Sun Han and M. Hoshi, Interval algorithm for random number generation, *IEEE Trans. Inf. Theory* **43**, 599 (1997).