Article:
Cizauskas, M., Tebyanian, H., Fox, M. et al. (2026) 33 Gbit/s source-device-independent quantum random number generator based on heterodyne detection with real-time FPGA-integrated extraction. Quantum Science and Technology, 11 (2). 025022. ISSN: 2058-9565

https://doi.org/10.1088/2058-9565/ae4d81

Quantum Science and Technology

PAPER • **OPEN ACCESS**

33 Gbit/s source-device-independent quantum random number generator based on heterodyne detection with real-time FPGA-integrated extraction

To cite this article: Marius Cizauskas *et al* 2026 *Quantum Sci. Technol.* **11** 025022

IOP Publishing Quantum Sci. Technol. 11 (2026) 025022 https://doi.org/10.1088/2058-9565/ae4d81

Quantum Science and Technology

OPEN ACCESS

RECEIVED
25 November 2025

REVISED
6 February 2026

ACCEPTED FOR PUBLICATION
4 March 2026

PUBLISHED
16 March 2026

PAPER

33 Gbit/s source-device-independent quantum random number generator based on heterodyne detection with real-time FPGA-integrated extraction

Marius Cizauskas[1,2,*], **Hamid Tebyanian**[3], **A Mark Fox**[2], **Manfred Bayer**[1], **Marc Assmann**[1] and **Alex Greilich**[1,*]

[1] Experimentelle Physik 2, Technische Universität Dortmund, 44221 Dortmund, Germany
[2] School of Mathematical and Physical Sciences, University of Sheffield, Sheffield, United Kingdom
[3] School of Physical and Chemical Sciences, Queen Mary University of London, London, United Kingdom
[*] Authors to whom any correspondence should be addressed.

E-mail: marius.cizauskas@tu-dortmund.de and alex.greilich@tu-dortmund.de

Keywords: QRNG, heterodyne, real-time, vacuum fluctuations, source-device-independence

Abstract

We present a high-speed continuous-variable quantum random number generator (QRNG) based on heterodyne detection of vacuum fluctuations. The scheme follows a source-device-independent security model in which the entropy originates from quantum measurement uncertainty and no model of the source is required; security depends only on the trusted measurement device and the calibrated discretization, and thus remains valid even under adversarial state preparation. The optical field is split by a 90 ° optical hybrid and measured by two balanced photodiodes to obtain both quadratures of the vacuum state simultaneously. The analog outputs are digitized using a dual-channel 12-bit analog-to-digital converter operating at a sampling rate of 3.2 GS s^{-1} per channel, and processed in real time by an field-programmable gate array (FPGA) implementing Toeplitz hashing for randomness extraction. The quantum-to-classical noise ratio was verified through calibrated power spectral density measurements and cross-checked in the time domain, confirming vacuum-noise dominance within the 1.6 GHz detection bandwidth. After extraction, the system achieves a sustained generation rate of $R_{\mathrm{net}} = 33.92$ Gbit s^{-1} of uniformly distributed random bits, which pass all NIST and Dieharder statistical tests. The demonstrated platform provides a compact, FPGA-based realization of a practical heterodyne continuous-variable source-independent QRNG suitable for high-rate quantum communication and secure key distribution systems.

1. Introduction

Random numbers serve as fundamental building blocks across numerous fields, from Monte Carlo simulations in scientific research [1] to quantum key distribution based cryptographic protocols [2] of which the purpose is to provide unbreakable secure encrypted communication [3]. In cryptographic applications, the quality and unpredictability of random numbers directly determine the security of encryption schemes, since they are used in one-time pad encryption, where the key length must match the message length and cannot be reused [2]. While classical random number generators based on deterministic algorithms can satisfy many computational needs, they fundamentally cannot provide the true unpredictability required for security-critical applications, as their outputs are ultimately predictable given sufficient computational resources and knowledge of the algorithm's internal state [4].

Quantum mechanics offers a fundamental solution to this limitation through the intrinsic randomness present in quantum measurements. quantum random number generators (QRNGs) exploit the fundamental uncertainty principle of quantum mechanics to produce truly unpredictable random numbers,

Table 1. Comparison of recent experimental QRNG demonstrations (2018–2025) focusing on security assumptions, achieved random bit rate, extraction mode, and entropy estimation method.

Work (Year)	Security Model	Bit Rate	Extraction	Entropy Generation Method
Avesani *et al* (2018) [10]	Source-device-independent	17.42 Gbps	Offline	Heterodyne vacuum noise measurements
Zheng *et al* (2019) [9]	Device-dependent	6 Gbps	Online (FPGA)	Homodyne vacuum noise measurements
Bruynsteen *et al* (2023) [11]	Device-dependent	100 Gbps	Offline	Homodyne vacuum noise measurements
Bertapelle *et al* (2023) [12]	Source-device-independent	20 Gbps	Offline	On-chip heterodyne vacuum noise
Qiu *et al* (2025) [13]	Source-device-independent	347 Mbps	Online (FPGA)	Fiber beamsplitter statistics, device imperfection tolerant
Cheng *et al* (2024) [14]	Semi-device-independent	580 Mbps	Offline	Entropy from squeezed light states
Zhang *et al* (2025) [15]	One-sided device-independent	7.06 Mbps	Offline	Entropy from quantum steering
Yang *et al* (2025) [16]	Device-dependent	156 Gbps	Offline	Laser phase noise
Crampton *et al* (2025) [17]	Device-dependent	2 Gbps	Online (FPGA)	Phase-diffusion in two gain-switched lasers
This work	Source-device-independent	33.92 Gbps	Online (FPGA)	Heterodyne vacuum noise measurements

providing a level of security that classical systems cannot achieve. The measurement of quantum observables yields outcomes that are fundamentally non-deterministic, even with complete knowledge of the quantum state prior to measurement. There are several established QRNG methods, such as measuring single photons going through a 50:50 beamsplitter [5], measurements of laser phase fluctuations [6], amplified spontaneous emission in fiber amplifiers [7], vacuum fluctuations [8, 9] and etc. QRNGs are further divided into two main categories. There are discrete variable QRNGs, where the random numbers are formed by quantum measurements that yield binary results, such as single-photon measurements, and there are continuous variable QRNGs, where continuous quantum mechanical properties are measured.

The critical distinction between online and offline extraction is operationally significant: offline schemes require intermediate data storage and batch post-processing, so the advertised rate represents a peak throughput that cannot be sustained on demand. Online extraction, by contrast, produces certified random bits continuously at line rate, enabling direct integration with high-speed cryptographic applications.

In recent times, QKD systems have been improving and are operating in the GHz range [18, 19]. With the increasing QKD operating frequency, higher bandwidth QRNGs are needed to match the data transmission rates. There are already implementations showing bandwidths of more than a hundred Gbits/s [11, 20]. However, these systems are not fully secure. The sources or the measurement devices can be tampered, which reduces the quality of QRNG, allowing for potential attacks [21]. Device-independent (DI) QRNGs represent the gold standard of security, assuming neither source nor measurement device is trusted and employing self-testing protocols to certify randomness. However, the loophole-free Bell tests required for DI certification impose severe experimental constraints, limiting demonstrated rates to the kbit s^{-1} regime [22]. Source-DI (SDI) QRNGs offer a practical intermediate: by trusting only the measurement device, they eliminate the need for Bell violations while retaining security against adversarial or drifting sources. Several SDI implementations have demonstrated rates in the 10–30 Gbit s^{-1} range [10, 23, 24].

However, a critical limitation persists: in all prior high-rate SDI demonstrations, randomness extraction is performed offline. Raw data must be stored, transferred to a computer, and batch-processed after acquisition. The advertised bit rates therefore represent peak throughput under ideal post-processing conditions, not sustained real-time output. For applications requiring random bits on demand, such as QKD systems operating at GHz clock rates, this offline bottleneck renders the quoted rates unachievable in practice. Closing this gap requires implementing the full extraction pipeline in real-time hardware. A comparison between the different works can be seen in table 1.

In this paper, we present a SDI QRNG based on heterodyne measurements of vacuum fluctuations, with acquisition, extraction and data transfer performed entirely on field-programmable gate

array(FPGA) in real time. While we adopt the SDI security framework established by [10], our security analysis extends it in several respects necessary for real-time operation: explicit treatment of analog-to-digital converter (ADC) clipping as trusted post-selection, conservative inflation factors for excess noise, temporal-mode definition addressing sample correlations, and security composition including calibration uncertainty. Together with the engineering implementation, this constitutes the first demonstration of SDI-certified random number generation with online extraction exceeding 30 Gb s^{-1}. By using an FPGA together with a dual-channel ADC (ADC12DJ5200RFEVM) operating at a sampling rate of $f_s = 3.2$ GHz per channel (up to 5.2 GHz maximum), both heterodyne detection channels are digitized in parallel to generate random numbers. A net extracted throughput of $R_{\text{net}} = 33.92$ Gbit s^{-1} is achieved on this setup. In the subsequent sections, we cover the experimental setup, including the FPGA implementation, the results of the QRNG, and its statistical tests.

2. Experimental details

2.1. General setup

The experimental setup, shown in figure 1, implements a heterodyne detection scheme for vacuum fluctuation-based quantum random number generation. The system is built on a continuous-wave laser source LPSC-1550-FC with CLD1010LP laser diode driver from Thorlabs operating at wavelength $\lambda = 1550$ nm with output powers up to 30 mW. The laser output serves as the local oscillator (LO). It passes through an electrically variable optical attenuator (EVOA) EVOA1550A to precisely control the LO power and to ensure its stability. The EVOA is set to limit the detector-plane LO power to $\leqslant 15$ mW for the variance–slope calibration range.

The LO is input to a 90 ° optical hybrid Optoplex model HB-C0AFAS066 operating in the telecom C-band. This device implements a 2×4 optical hybrid coupler that mixes the LO E_{LO} with the reference signal (vacuum in this case) E_{vac} from the environment to generate four quadratural states in the complex-field space.

Each of the four hybrid outputs is individually controlled by EVOAs Thorlabs V1550A. These devices enable real-time gain balancing which allows for 90 ° optical hybrid outputs to be properly balanced. This ensures equal power distribution to maintain the quadrature measurement accuracy and compensate for any asymmetries in the optical path.

The four EVOA outputs are paired and fed into two Thorlabs PDB480C-AC balanced photodiode detectors (BPDs). Each BPD consists of a pair of matched photodiodes operated in differential mode. The 90 ° optical hybrid routes the LO and the signal (vacuum at the signal port) so that ports (1,2) and (3,4) form two balanced pairs with a relative phase shift of 180 ° within each pair and 90 ° between the two pairs. With the diodes in differential mode, these two balanced differences implement a simultaneous measurement of orthogonal quadratures (heterodyne). Denoting the differential voltages by A and B (for the two BPDs), we model

$$\begin{aligned} A \equiv V_X &= \kappa_X \sqrt{P_{\text{LO}}^{(\text{det})}}\, X + n_{\text{e},X}, \\ B \equiv V_P &= \kappa_P \sqrt{P_{\text{LO}}^{(\text{det})}}\, P + n_{\text{e},P}, \end{aligned} \tag{1}$$

where $\kappa_{X/P}$ are responsivities [$\text{V}\sqrt{\text{W}}$], X, P are heterodyne quadrature outcomes with $\langle X \rangle = \langle P \rangle = 0$ and $\text{Var}(X) = \text{Var}(P) = \frac{1}{2}$, and $n_{\text{e},X/P}$ are electronics noises (V).

Hence the shot-noise slopes satisfy

$$m_X = \frac{\kappa_X^2}{2}, \qquad m_P = \frac{\kappa_P^2}{2}. \tag{2}$$

Consequently, the variances scale linearly with LO power,

$$\begin{aligned} \text{Var}\,(A) &= m_X P_{\text{LO}}^{(\text{det})} + \sigma_{\text{e},X}^2, \\ \text{Var}\,(B) &= m_P P_{\text{LO}}^{(\text{det})} + \sigma_{\text{e},P}^2, \end{aligned} \tag{3}$$

where $\sigma_{\text{e},X}^2$, $\sigma_{\text{e},P}^2$ are the classical electronics noise variances (amplifier thermal noise, photodiode dark current, ADC quantization), measured with the LO blocked. These correspond to the intercepts $c_X = \sigma_{\text{e},X}^2$ and $c_P = \sigma_{\text{e},P}^2$ in the slope–offset calibration of section 2.3.

The analog electrical signals from both BPDs, representing the X and P quadratures, first go through a ZX60-P105LN+ 15 dB amplifier to ensure that most of the ADC input range is covered, then they go through a series connected low-pass filter (LPF) and high-pass filter (HPF). The LPF is a SLP-1650+

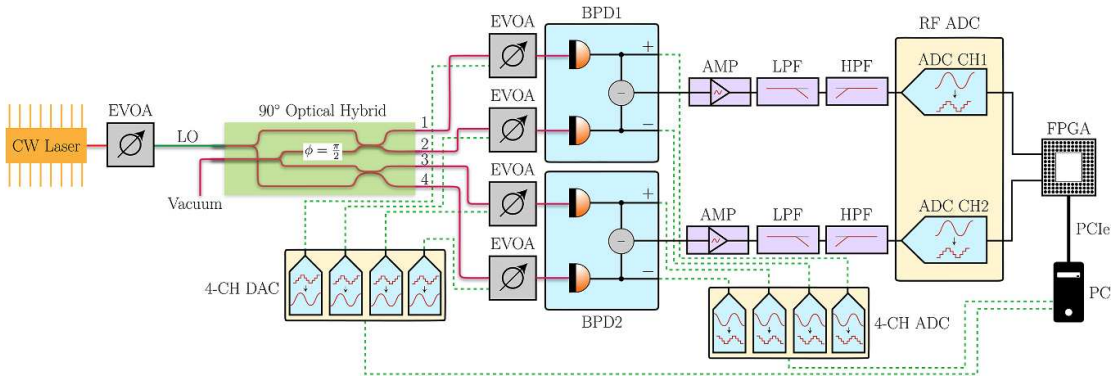

Figure 1. Schematic diagram of the vacuum fluctuation-based quantum random number generator using heterodyne detection. The setup consists of a laser source followed by an electrically variable optical attenuator (EVOA) for power control and stabilization. The laser output is mixed with vacuum fluctuations in a 90° optical hybrid, which simultaneously measures both X and P quadratures of the electromagnetic field. Each of the four hybrid outputs are individually controlled by EVOAs with electrical control inputs before being detected by balanced photodiode detectors (BPD1 and BPD2). The analog electrical signals from both detectors are first filtered by a low-pass and high-pass filters to filter out as much classical electrical noise as possible and then digitized by a 2-channel analog-to-digital converter (ADC) and processed in real-time by a field-programmable gate array (FPGA) connected via an FMC (FPGA Mezzanine Card) interface.

with a bandwidth of DC-1650 MHz and the HPF is a SHP-200+ with a bandwidth of 185–3000 MHz, yielding an effective bandpass filter with a bandwidth of 185–1650 MHz. The upper limit was chosen at 1650 MHz since the diodes have a bandwidth of up to 1600 MHz and the sampling rate is 3.2 GHz, which puts the bandwidth slightly outside the Nyquist frequency; the *LPF* nominal passband is DC-1400 MHz with a -3 dB point at 1650 MHz. The lower bound was chosen at 185 MHz to reduce the classical lower frequency noise originating in the diodes. Lastly, the diode outputs are digitized by a 2-channel ADC ADC12DJ5200RFEVM with a sampling rate of $f_s = 3.2$ GHz set for both channels and a resolution of 12 bits. The digitized quadrature data streams are transmitted to a FPGA Virtex Ultrascale+ VCU118 development board via a high-speed FMC (FPGA Mezzanine Card) interface. The FPGA implements a real-time randomness extraction algorithm via Toeplitz hashing to generate provably random bit sequences from the quantum vacuum fluctuations by flattening the Gaussian distribution of the data.

2.2. Randomness generation and extraction

Security model

We model the QRNG as a prepare–measure protocol on n_{eff} temporal modes $A^{n_{\text{eff}}}$ impinging on a trusted receiver. The optical input is treated as fully untrusted: an adversary may prepare an arbitrary joint state $\rho_{A^{n_{\text{eff}}}E}$ of the incoming modes and a quantum memory E, allowing arbitrary entanglement and arbitrary inter-round correlations. The receiver is trusted and fixed prior to acquisition: the 90 ° hybrid, balanced detection, analogue front-end, ADC digitization, the discretization map (binning and the clipping rule), and the extractor circuit are treated as trusted. The Toeplitz seed is assumed uniform and independent of the measured data; Toeplitz hashing is a strong (universal$_2$) extractor.

In the SDI setting adopted here, security is enforced by the trusted receiver measurement rather than by a model of the optical source. The trusted receiver implements a discretized heterodyne POVM $\{M_z\}$ whose elements are determined by the calibrated discretization in phase space; the resulting entropy bound depends only on these trusted bin regions through operator-norm bounds on $\|M_z\|_\infty$.

Minimum entropy quantifies the worst-case unpredictability of a random variable in the presence of an adversary. For a random variable X with probability distribution $p(x)$,

$$H_{\min}(X) = -\log_2\left(\max_x p(x)\right). \tag{4}$$

In QRNG security we use the smooth conditional min-entropy $H_{\min}^{\epsilon_s}(\cdot|E)$ where E denotes adversary's (quantum) side information; we write H for classical protocol history when needed. Let Z denote the discretized heterodyne outcome (ADC bin index). For a classical outcome Z correlated with quantum side information E, the (non-smooth) conditional min-entropy is defined operationally by the adversary's optimal guessing probability

$$\begin{aligned} H_{\min}(Z|E) &= -\log_2 P_{\text{guess}}(Z|E), \\ P_{\text{guess}}(Z|E) &= \max_{\{E_z\}} \sum_z \text{Tr}[(M_z \otimes E_z)\,\rho_{AE}], \end{aligned} \tag{5}$$

where $\{M_z\}$ is the trusted POVM implemented on the optical mode and $\{E_z\}$ is an arbitrary POVM on E. Defining

$$c := \max_z \|M_z\|_\infty, \tag{6}$$

the inequality $M_z \leqslant \|M_z\|_\infty \mathbb{I} \leqslant c\mathbb{I}$ implies, for any ρ_{AE},

$$\sum_z \text{Tr}[(M_z \otimes E_z)\,\rho_{AE}] \leqslant \text{Tr}\left[\left(c\mathbb{I} \otimes \sum_z E_z\right)\rho_{AE}\right] = c, \tag{7}$$

and hence $P_{\text{guess}}(Z|E) \leqslant c$ and $H_{\min}(Z|E) \geqslant -\log_2 c$. For discretized heterodyne, lemma 1 bounds $\|M_z\|_\infty$ in terms of the trusted bin area, yielding the state-independent bound used below.

Smooth conditional min-entropy evaluates non-smooth entropies on the optimal state within an ϵ-neighborhood of the measured state, where ϵ is the security parameter, while the per-round heterodyne discretization bound is state-independent and holds *unsmoothed* [10]:

$$H_{\min}(X|E) \geqslant -\log_2\left(\min\left\{\tfrac{\delta_X'\delta_P'}{2\pi}, 1\right\}\right). \tag{8}$$

The heterodyne POVM is $\Pi(\alpha) = |\alpha\rangle\langle\alpha|/\pi$, so $Q_\rho(\alpha) = \langle\alpha|\rho|\alpha\rangle/\pi$ and $\sup_\rho Q_\rho = 1/\pi$. With $\text{Var}(X) = \text{Var}(P) = \frac{1}{2}$ we have $\alpha = (X + iP)/\sqrt{2}$ and $\text{d}^2\alpha = \text{d}X\text{d}P/2$. If the ADC defines rectangles $\{R_{mn}\}$ of area $\delta_X'\delta_P'$ in (X, P), then $\text{Area}_\alpha(R_{mn}) = \delta_X'\delta_P'/2$. Hence (using $\text{d}^2\alpha = \text{d}X\text{d}P/2$ for vacuum units $\text{Var}(X) = \text{Var}(P) = 1/2$)

$$p_{\max} := \max_{m,n} \int_{R_{mn}} Q_\rho(\alpha)\,\text{d}^2\alpha \;\leqslant\; \min\left\{\tfrac{\delta_X'\delta_P'}{2\pi}, 1\right\}, \tag{9}$$

and

$$h_{\min}^{(1)} \;\geqslant\; -\log_2\left(\min\left\{\tfrac{\delta_X'\delta_P'}{2\pi}, 1\right\}\right). \tag{10}$$

For any (possibly adversarial) preparation and any past transcript,

$$\max_{m,n} \Pr[(X,P) \in R_{mn} \,|\, \text{history}, \mathcal{E}] \leqslant \min\left\{\tfrac{\delta_X'\delta_P'}{2\pi}, 1\right\}, \tag{11}$$

and thus $h_{\min}^{(1)} \geqslant -\log_2\left(\min\{\frac{\delta_X'\delta_P'}{2\pi}, 1\}\right)$ holds per round without IID assumptions and is adaptivity-robust.

Lemma 1. *Let $M_R = \int_R |\alpha\rangle\langle\alpha| \frac{\text{d}^2\alpha}{\pi}$ be the heterodyne POVM element for a measurable bin $R \subset \mathbb{C}$. Then*

$$\|M_R\|_\infty \;\leqslant\; \min\left\{\tfrac{\text{Area}(R)}{\pi}, 1\right\}. \tag{12}$$

Proof. For any unit vector $|\psi\rangle$, $\langle\psi|M_R|\psi\rangle = \int_R Q_\psi(\alpha)\,\text{d}^2\alpha$ with $Q_\psi(\alpha) = |\langle\psi|\alpha\rangle|^2/\pi \leqslant 1/\pi$; hence $\langle\psi|M_R|\psi\rangle \leqslant \text{Area}(R)/\pi$. Completeness of the POVM gives $\int_{\mathbb{C}} |\alpha\rangle\langle\alpha| \frac{\text{d}^2\alpha}{\pi} = \mathbb{I}$, so $0 \leqslant M_R \leqslant \mathbb{I}$ and $\|M_R\|_\infty \leqslant 1$. Taking the tighter of the two bounds yields the stated minimum. □

Boundary (rail-adjacent but non-clipped) bins are treated as truncated rectangles with their actual area used in (9). ADC *clipping* codes result from voltages outside the input range $[-V_{\text{pp}}/2, V_{\text{pp}}/2]$ of the ADC. They correspond to semi-infinite regions and are excluded. Dropping clips is a post-selection on the event $\Omega_{\text{acc}} := \{\text{no clip}\}$ determined solely by the trusted ADC range. For security we upper bound the rejection probability by a conservative value $f_{\text{clip}}^{\max}$ fixed by the trusted ADC range and operating point (with any estimation uncertainty absorbed into β), so that $\Pr[\Omega_{\text{acc}}] \geqslant 1 - f_{\text{clip}}^{\max}$. Consequently, the single-round bound for the post-selected data becomes

$$h_{\min,\text{no clip}}^{(1)} \;\geqslant\; -\log_2\left(\min\left\{\tfrac{\delta_X'\delta_P'}{2\pi\left(1-f_{\text{clip}}^{\max}\right)}, 1\right\}\right). \tag{13}$$

and the extraction length obeys

$$\ell \leqslant n_{\text{valid}}\, h^{(1)}_{\text{min, no clip}} - 2\log_2 \frac{1}{\varepsilon_{\text{PA}}}, \qquad n_{\text{valid}} = n\left(1 - f_{\text{clip}}\right), \tag{14}$$

where n is the total number of rounds entering extraction.

2.3. Calibration and vacuum-unit resolutions

The bound depends only on the discretization induced by the apparatus. We determine the raw resolutions $\delta_{X/P}$ and the conservative resolutions $\delta'_{X/P}$ from a slope–offset calibration and the ADC's effective code width as follows.

$$\sigma^2_{V,X}(P_{\text{LO}}) = m_X P_{\text{LO}} + c_X, \qquad \sigma^2_{V,P}(P_{\text{LO}}) = m_P P_{\text{LO}} + c_P, \tag{15}$$

with $m_{X/P}$ in V^2/W and $c_{X/P} \geqslant 0$ in V^2. To exclude LO relative-intensity-noise leakage due to residual imbalance, we fit

$$\sigma^2_{V,X}(P_{\text{LO}}) = m_X P_{\text{LO}} + c_X + \eta_X P^2_{\text{LO}}, \qquad \sigma^2_{V,P}(P_{\text{LO}}) = m_P P_{\text{LO}} + c_P + \eta_P P^2_{\text{LO}}, \tag{16}$$

and test $\eta_{X/P} = 0$ via lack-of-fit. We found $|\eta_{X/P}| P_{\text{LO,max}} \ll m_{X/P}$ within the calibration range; thus linear shot-noise slopes are valid. If $\eta \neq 0$, its effect is absorbed by increasing $\gamma_{X/P}$ in (22).

Here $c_X = \sigma^2_{\text{e},X}$ and $c_P = \sigma^2_{\text{e},P}$, consistent with the detector model in (1). Since $\text{Var}(X) = \text{Var}(P) = \frac{1}{2}$ for vacuum, the conversion factors to vacuum units are:

$$\alpha_X = \sqrt{2 m_X P^{(\text{det})}_{\text{LO}}}, \quad \alpha_P = \sqrt{2 m_P P^{(\text{det})}_{\text{LO}}}, \tag{17}$$

here $P^{(\text{det})}_{\text{LO}}$ denotes the *per-BPD* LO power at the photodiodes *after* the 90 ° hybrid and EVOA balancing (i.e. the sum incident on the two diodes of a given BPD). All power readings used in (17) are taken at this plane. We define per-sample vacuum-unit coordinates by

$$\hat{X} := \frac{A}{\alpha_X}, \qquad \hat{P} := \frac{B}{\alpha_P}, \tag{18}$$

so that for vacuum $\text{Var}(\hat{X}) = \text{Var}(\hat{P}) = \frac{1}{2}$. With ADC peak-to-peak range $V_{\text{pp}} = 0.5\,\text{V}$ after the analog front end, define the effective code width

$$\Delta V_{\text{eff}} = \frac{V_{\text{pp}}}{2^{n_{\text{ENOB}}}}, \tag{19}$$

where $n_{\text{ENOB}} = 10.2$ (X channel) and 10.3 (P channel) is the measured effective number of bits of the ADC (sine-wave histogram method, averaged over 185–1600 MHz). Datasheet ENOB variation is absorbed into β. The raw resolutions are

$$\delta_X = \frac{\Delta V_{\text{eff}}}{\alpha_X}, \qquad \delta_P = \frac{\Delta V_{\text{eff}}}{\alpha_P}. \tag{20}$$

Let σ^2_X and σ^2_P denote the measured quadrature variances expressed in vacuum units after applying the gains $\alpha_{X/P}$ from (17). To conservatively include excess noise or calibration uncertainty, define

$$\gamma_X := \max\left\{1, \sqrt{\sigma^2_X / (1/2)}\right\}, \qquad \gamma_P := \max\left\{1, \sqrt{\sigma^2_P / (1/2)}\right\}, \tag{21}$$

$$\delta'_X = \gamma_X \delta_X, \qquad \delta'_P = \gamma_P \delta_P. \tag{22}$$

More generally, the single-shot bound reduction is $\log_2(\gamma_X \gamma_P)$. With the measured (σ^2_X, σ^2_P), the single-shot penalty is

$$\begin{aligned}\Delta h_{\min}^{(1)} &= \log_2(\gamma_X\gamma_P)\\ &= \tfrac{1}{2}\log_2\left(\max\left\{1,2\sigma_X^2\right\}\right)+\tfrac{1}{2}\log_2\left(\max\left\{1,2\sigma_P^2\right\}\right).\end{aligned} \tag{23}$$

For $(\sigma_X^2,\sigma_P^2)=(0.616,0.647)$ this gives $\Delta h_{\min}^{(1)}\approx 0.336$ bits/round. In the symmetric case $\sigma_X^2=\sigma_P^2=\frac{1}{2}(1+\xi)$, where ξ is the relative excess noise above the vacuum variance, one has $\gamma_X=\gamma_P=\sqrt{1+\xi}$ and the reduction equals $\log_2(1+\xi)$. The guard $\gamma\geqslant 1$ prevents artificial entropy gain if $\sigma^2<1/2$ due to calibration uncertainty. Imperfections in the 90° optical hybrid and detection chain (non-ideal phase shift, unequal splitting ratios, detector gain mismatch) can induce correlation between the X and P quadratures. We measured the raw cross-quadrature correlation to be $|\rho_{XP}|=0.012\pm 0.004$ at zero lag. To verify no correlation structure exists at nonzero delays, we computed the full cross-correlation function as a function of sample lag (figure 8); the cross-correlation remains below 10^{-3} at all measured delays, confirming that coupling is confined to the zero-lag term and does not introduce temporal structure.

This correlation does not compromise the security bound. The per-round bound $p_{\max}\leqslant \delta_X'\delta_P'/(2\pi)$ derives from the operator norm of the POVM element (lemma 1), which depends only on the bin area in phase space, not on the correlation structure of the input state. An adversary could prepare an arbitrarily correlated state, and the bound would still hold because it is a property of the trusted measurement, not the source. At our measured $|\rho_{XP}|=0.012$, a hypothetical correlation penalty would be $\Delta h_{\text{corr}}=-\frac{1}{2}\log_2(1-\rho_{XP}^2)<10^{-4}$ bits/round, which is negligible. No decorrelation step is required.

We propagate uncertainties in $(m_{X/P},c_{X/P},V_{\text{pp}},n_{\text{ENOB}})$ to (δ_X',δ_P') and to $h_{\min}^{(1)}$ by first-order error propagation; error bars for $\delta_{X/P}'$ and the net rate include these calibration contributions.

The security proof relies on the fact that for any quantum state, the Husimi function $Q_\rho(q+ip)$ is upper bounded by $1/\pi$. This fundamental quantum mechanical constraint ensures that even if an adversary prepares an optimal quantum state to maximize their guessing probability, the conditional min-entropy remains bounded by the measurement resolution alone [10]. This bound is independent of the actual quantum state prepared by the potentially malicious source, providing information-theoretic security guarantees. Practical implementations must account for device imperfections that can compromise this security. Imbalanced heterodyne detection, arising from non-ideal beam splitter ratios or photodiode efficiency mismatches, introduces LO fluctuation that contributes excess noise. Under imbalanced conditions, LO fluctuation cannot be eliminated and affects the calibration of vacuum fluctuations [25]. Without proper consideration of these fluctuations, the extractable randomness can be overestimated, creating security vulnerabilities. For this reason, we use additional EVOAs at the 90 ° optical hybrid, since it does not show perfect 50:50 splitting ratio.

As mentioned before, upon initial measurements of this system, the resulting distribution is Gaussian, which makes it easier to predict the generated random numbers. For this reason, a randomness extractor has to be used to flatten the distribution and extract the QRNG which is mixed with classical noise sources. One of the more commonly used extractors is Toeplitz hashing [8–11, 24, 26], belonging to universal$_2$ family and hence a *strong* extractor [27, 28]: for a public, uniform, independent seed S, $(S,\text{Ext}(Z^n,S))\approx(S,U_j)$.

We reuse the same public seed S across all parallel output blocks within one extractor invocation; Toeplitz hashing is a strong (universal$_2$) extractor, and our single-round heterodyne min-entropy bound is independent of S and the classical transcript. Hence, by a hybrid/triangle-inequality argument applied block by block, the joint trace distance satisfies

$$\left\|\left(S,\text{Ext}\left(Z_{(1)}^n,S\right),\ldots,\text{Ext}\left(Z_{(B)}^n,S\right),E\right)-\left(S,U_{\ell_1},\ldots,U_{\ell_B},E\right)\right\|_{\text{tr}}\leqslant\sum_{i=1}^{B}\left(\varepsilon_{\text{s}}^{(i)}+\varepsilon_{\text{PA}}^{(i)}\right) \tag{24}$$

where $\varepsilon_{\text{s}}^{(i)}$ is the smoothing error of a block and $\varepsilon_{\text{PA}}^{(i)}$ is the privacy amplification error of a block. In this way we allocate $(\varepsilon_{\text{s}}^{(i)},\varepsilon_{\text{PA}}^{(i)})$ so that the sum over blocks equals the advertised ε_{sec}.

The two 12-bit ADC codes are concatenated into a single 24-bit symbol $D_i:=(X_i\|P_i)$ at each sampling instant. This concatenation is mandated by the security model: the heterodyne POVM $\Pi(\alpha)=|\alpha\rangle\langle\alpha|/\pi$ is a joint measurement on 2D phase space that cannot be factored into independent marginal measurements. The min-entropy bound (equation (10)) certifies entropy of the pair (X,P) via the joint bin area; treating the quadratures as separate streams would require a different security analysis based on marginal distributions rather than the joint Husimi bound.

The heterodyne bound is state- and history-independent: for every retained round $t\leqslant n_{\text{eff}}$ and any classical transcript H,

$$p_{\max}^{(t)}\leqslant\min\left\{\tfrac{\delta_X'\delta_P'}{2\pi},1\right\}. \tag{25}$$

Because the trusted device applies the same per-round POVM $\{M_z\}$ to each temporal mode and does not adapt to H, one has for any cq state with arbitrary inter-round correlations

$$\begin{aligned} H_{\min}^{\varepsilon_s}(Z^{n_{\text{eff}}}|E,H) \;&\geqslant\; n_{\text{eff}}\, h_{\min}^{(1)}, \\ h_{\min}^{(1)} &= -\log_2\left(\min\left\{\tfrac{\delta_X'\delta_P'}{2\pi}, 1\right\}\right). \end{aligned} \tag{26}$$

where n_{eff} is the number of temporal modes on which the trusted POVM $\{M_z\}$ acts identically.

We report integrated autocorrelation time τ_{int} as a diagnostic of classical correlations in the sampled voltage stream. A conservative way to enforce approximate factorization is to define rounds on a decimated stream with spacing D samples (e.g. $D \geqslant \lceil \tau_{\text{int}} \rceil$), leading to $n_{\text{eff}} = n/D$ and an effective sampling rate $f_s^{(\text{eff})} = f_s/D$. In our reported security analysis and rates, we instead apply a fixed unit-Jacobian orthonormal transform on blocks before discretization, so that the n samples are mapped to n orthonormal temporal modes measured by the same POVM $\{M_z\}$; in this case we take $n_{\text{eff}} = n$ and $f_s^{(\text{eff})} = f_s$, and τ_{int} is used purely as a diagnostic check that residual correlations after extraction are within statistical error. The analog front-end and sampling chain are trusted and absorbed into the calibration that defines $\delta'_{X/P}$. After enforcing factorization as above, the product-POVM structure holds on the retained temporal modes and we use $H_{\min}(Z^{n_{\text{eff}}}|E) \geqslant n_{\text{eff}} h_{\min}^{(1)}$.

We separate the secrecy parameter from statistical-estimation error, thus we define

$$\varepsilon_{\text{sec}} \;=\; \varepsilon_s + \varepsilon_{\text{PA}} + \beta, \tag{27}$$

where β upper-bounds the probability that calibration parameters deviate outside their quoted confidence intervals. Privacy amplification obeys

$$\ell \leqslant H_{\min}^{\varepsilon_s}(Z^{n_{\text{eff}}}|E) - 2\log_2 \frac{1}{\varepsilon_{\text{PA}}}. \tag{28}$$

Here E denotes adversarial (quantum) side information. Since $H_{\min}^{\varepsilon_s}(\cdot|E) \geqslant H_{\min}(\cdot|E)$ for any $\varepsilon_s \in [0,1)$, our unsmoothed lower bound $H_{\min}(Z^{n_{\text{eff}}}|E) \geqslant n_{\text{eff}} h_{\min}^{(1)}$ also lower-bounds the smoothed quantity used in privacy amplification.

Here β upper-bounds the probability that the calibration parameters $(m_{X/P}, c_{X/P}, \Delta V_{\text{eff}}, n_{\text{ENOB}})$ deviate outside their quoted confidence intervals; it is set via conservative concentration bounds for the variance fits and datasheet limits. We set $\varepsilon_s = \varepsilon_{\text{PA}} = \beta = 2^{-64}$, yielding $\varepsilon_{\text{sec}} = 3 \cdot 2^{-64}$. If the data are output in N_{blk} extractor invocations (fresh or reused public seed), we allocate per-block parameters $(\varepsilon_s/N_{\text{blk}}, \varepsilon_{\text{PA}}/N_{\text{blk}}, \beta/N_{\text{blk}})$ so that a union bound guarantees the advertised global bound on ε_{sec}.

2.4. FPGA implementation

The FPGA receives 20 samples for each channel per 160 MHz clock cycle, or 480-bits of data, over 16 JESD204C data lanes. This results in a total data rate of 76.8 Gbps at 3.2 GHz sample rate and 2-channel acquisition. Therefore, both pipelining and parallelization were applied to ensure that the data is processed in real-time. Figure 2 shows how the Toeplitz extraction scheme was implemented. In the figure shown, there are 5 pipeline stages; however, there is an additional initial stage for buffering ADC input, which is covered in further text.

Once there is data available in the input buffer, the pipeline starts with the Toeplitz generation stage. It can be seen that there is an OR gate for two signals, one which is triggered when Toeplitz extraction is done and another called Toeplitz init, which is only set when the FPGA is reset, so as to ensure proper Toeplitz matrix initialization on FPGA power on. If one of these signals is triggered, the initial Toeplitz vector signal T is set from the independently generated seed. When neither of the two aforementioned signals is triggered, it indicates that Toeplitz extraction is running, therefore T is shifted to the right by the combined bit-depth b. The right shift is needed because a block of the Toeplitz matrix T_B is taken in the next pipeline data loading stage. A total of $j+b$ bits are taken for a block, which effectively corresponds to b columns of the Toeplitz matrix, since, if the column size is j, the next column can be formed by $T_B(j+1:1)$ (T_B here denotes a block of the Toeplitz matrix). When T is shifted to the right by b, it results in taking further b columns of the Toeplitz matrix. In the same block, the b-bit ADC word D_i is taken every cycle from the buffer D.

Following is the multiplication stage, where first, the multiplication is done for the block of Toeplitz matrix, which is then stored in an intermediate two-dimensional vector U_B. As mentioned before, the numbers in the matrices are in binary, therefore mod 2 arithmetic applies. Therefore, the multiplication is calculated by applying AND bitwise operation on a single column of the Toeplitz matrix by a single

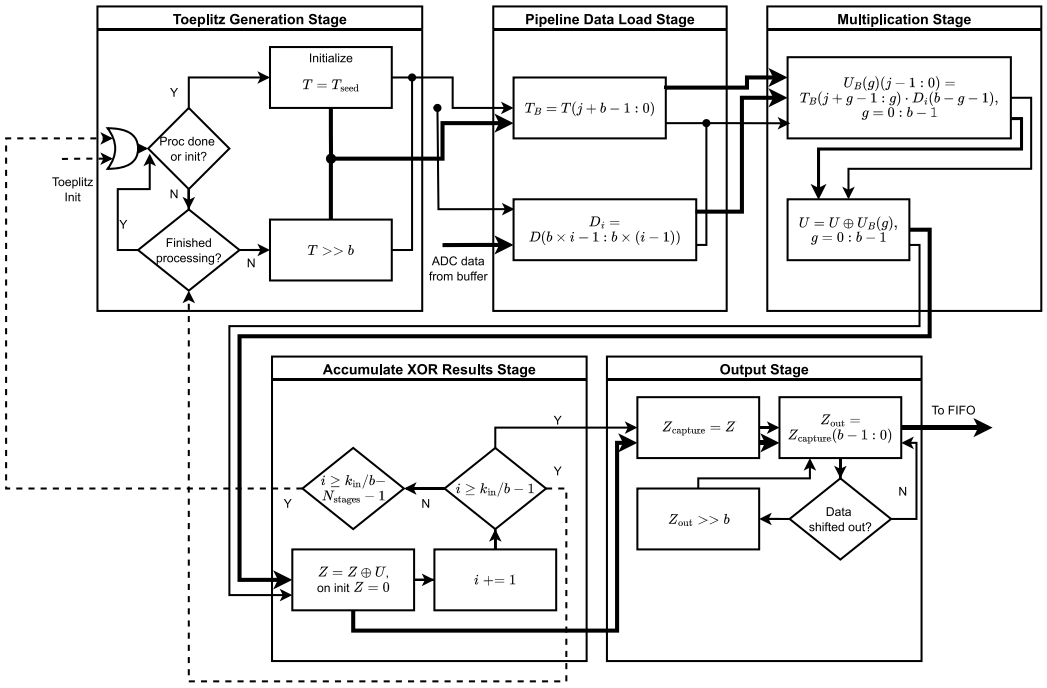

Figure 2. The pipelined Toeplitz extraction scheme. The scheme consists of separate pipeline stages. The dashed arrows indicate binary signals, the normal lines indicate the flow of the program and the bold, larger lines indicate flow of data. The square blocks represent operations and the diamond blocks represent conditionals.

bit of the ADC data D_i. This operation is done over a range of 0 to $b-1$ so as to multiply all b columns in the same cycle. Afterwards, a XOR reduction is done on U_{B}, which corresponds to the b-column multiplication sum, and is stored in U. After this, U goes to the XOR accumulation stage, where each clock cycle U is accumulated in the final result vector Z. This is done k_{in}/b times, until the ADC input buffer is fully processed. In this stage there are also two conditions: $i \geqslant k_{\text{in}}/b - 1$ is used to signal when Z is fully accumulated and it can be output, and $i \geqslant k_{\text{in}}/b - N_{\text{stages}} - 1$, where N_{stages} is the number of stages between the Toeplitz generation and accumulation stages (in this case 3), which is necessary to ensure timely initialization of T, since the Toeplitz generation stage operates ahead of the accumulation stage due to pipelining created latency.

When the $i \geqslant k_{\text{in}}/b - 1$ condition is satisfied, the output stage is started, where firstly Z is captured to Z_{capture} to ensure that the extracted value is not overridden by further calculations in the pipeline. The data is output in pieces of b bits for j/b cycles by shifting Z_{out} to the right by b. Lastly, it should be mentioned that between all these stages there's a single clock cycle latency and that initially all stages are inactive until the first Toeplitz stage becomes active.

Figure 2 shows only the extraction diagram of a single block. If f_{ADC} is the ADC sampling frequency (per channel), the total input rate is bf_{ADC} bits/s, where $b = 2n_{\text{ADC}} = 24$ bits/round. The FPGA receives N_{S} samples per cycle per channel, so the reference clock is $f_{\text{ref}} = f_{\text{ADC}}/N_{\text{S}}$. A single block processes b bits per clock cycle, i.e. bf_{ref} bits/s. Thus the number of parallel blocks required for full throughput is $N_{\text{B}} = \frac{bf_{\text{ADC}}}{bf_{\text{ref}}} = N_{\text{S}}$. In our setup $N_{\text{S}} = 20$, so 20 blocks are used to extract randomness in real time.

Figure 3 shows the full implementation of the FPGA design. Starting on the left, there is the JESD204C protocol output of the ADC, which goes into the transport layer, where first JESD204C lanes are appropriately mapped, then samples are extracted from the frames and finally the values are packed into 480-bit AXI4 Stream buses, where both CH1 and CH2 samples are interleaved in 24-bit words. This data is then sent to all the extraction blocks. All the blocks are interconnected by the buffering logic. The buffer is an additional pipeline stage part of the extraction schematic shown in figure 2. It consists of additional logic that ensures that the right amount of data is being read and to allow for proper operation with other parallel extraction blocks. More specifically, the buffer block additionally has a read enable input and a buffer full output. The buffer does not take any data until read enable is asserted and the buffer full output is asserted when the buffer is filled.

Such a design allows for daisy chaining the blocks and allow them to work in parallel. As seen in figure 3 the read enables and the buffer full signals are chained together, with the exception of the first

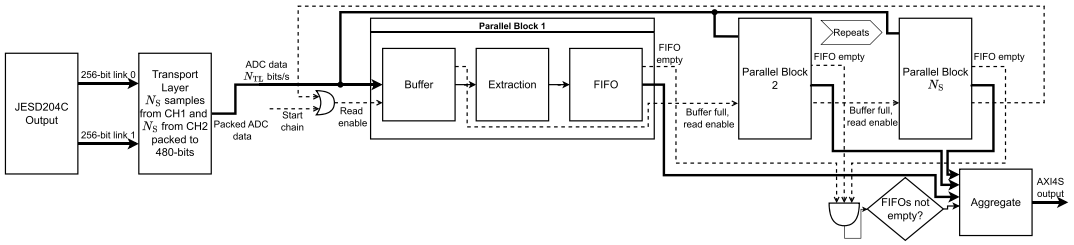

Figure 3. The overall schematic of the FPGA design. All parallel blocks contain the same functions and I/O, only the first block is fully drawn to make the diagram more compact. The dashed arrows indicate binary signals, the normal lines indicate the flow of the program and the bold, larger lines indicate flow of data.

Table 2. FPGA resource utilization for $j = 1272$ and $k_{\text{in}} = 2880$ Toeplitz matrix extraction algorithm with 20 parallel blocks.

Resource	Used	Available	Utilization (%)
LUT	596 422	1182 240	50.4
LUTRAM	12 697	591 840	2.1
FF	329 457	2364 480	13.9
BRAM	197	2160	9.1
DSP	20	6840	0.3
IO	24	832	2.9
GT	32	52	61.5
BUFG	25	1800	1.4
MMCM	1	30	3.3
PCIe	1	6	16.7

block, where there is an OR gate to the input of read enable. Initially, all of the blocks are inactive. Once ADC data becomes valid, the start chain signal connected to the OR gate of the first block read enable is asserted, which starts the ADC sample collection in the first block's buffer. Once the buffer is full, the buffer full signal is asserted and the collection starts in the second block and continues until the last block, where the buffer full signal is fed back to the first one, repeating the chain again. Finally, after extraction, the random bits are sent into a FIFO. The FIFO empty signals from each block are checked and when the condition of all FIFOs not being empty is satisfied, the FIFO outputs are read and are collated into a single vector signal.

To realize the full potential of the high bandwidth QRNG, the extracted data is sent over x16 PCIe to a computer. The PCIe interface on the FPGA runs with a clock speed of 250 MHz and 512-bit depth. To ensure proper clock domain crossing, the QRNG data first goes into an asynchronous FIFO, from which the PCIe interface reads. PCIe on the FPGA is handled by the XDMA IP provided by AMD. The computer that the FPGA is connected to runs Linux and uses the AMD provided XDMA driver to stream the data from the FPGA to the computer over an AXI4 Stream interface. There is an additional, not shown channel of the PCIe used to also transfer the raw, unextracted data to the computer. It is used both for security analysis and for monitoring the system during operation. The full implementation was successfully placed and routed without any timing issues. The used resources for a Toeplitz extraction for $j = 1272$, $k = 2880$ and 20 parallel blocks can be seen in table 2. The most significant usage is seen in the LUTs, which is expected, since even though the operations are not applied on a full matrix, the vector widths are still relatively large and therefore a lot of LUTs are needed to perform the bitwise operations over 20 independent extraction blocks. The matrix size could be slightly increased, however, at higher resource usage values the design has higher likelihood to fail implementation, since when usage is over 75%, congestion becomes an issue, making the placement and routing of logic challenging [29].

3. Results and discussion

3.1. Security analysis

The SDI security of our QRNG derives from the measurement-device characterization, not from statistical properties of the output. Specifically, the per-round min-entropy bound (equation (10)) follows

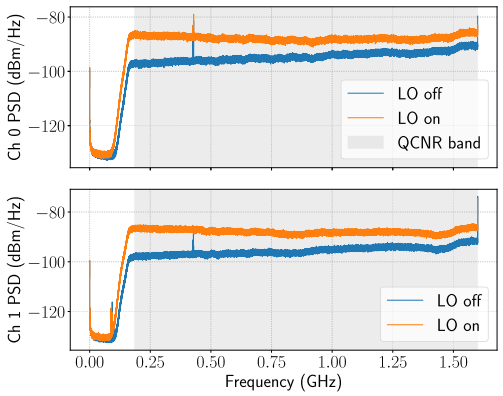

Figure 4. PSD of both ADC channels that are connected to the balanced diodes. The blue curve corresponds to no laser power and the orange curve to a local-oscillator power of 8.4 mW. The gray colored area shows over which range the PSD is integrated for QCNR.

from the operator norm of heterodyne POVM elements (lemma 1), which holds for *any* input state—including states prepared adversarially to maximize the guessing probability. The security-critical parameters are therefore the calibrated bin dimensions δ'_X, δ'_P and the privacy amplification composition $\varepsilon_{\text{sec}} = \varepsilon_{\text{s}} + \varepsilon_{\text{PA}} + \beta$.

In figure 4 we show the power spectral density (PSD) of both ADC channels when measuring with LO and without LO. The results are similar to other heterodyne and homodyne set-ups, where we see an LO-induced noise-power increase of $\approx 7.5\,\text{dB}$ (X) and $\approx 6.5\,\text{dB}$ (P) [9, 23, 30], indicating that quantum noise is present, not only the classical noise. In particular,

$$\begin{aligned} \Delta_{\text{dB}}^{(X)} &= 10\log_{10}\left(\frac{m_X P_{\text{LO}}^{(\text{det})} + c_X}{c_X}\right), \\ \Delta_{\text{dB}}^{(P)} &= 10\log_{10}\left(\frac{m_P P_{\text{LO}}^{(\text{det})} + c_P}{c_P}\right), \end{aligned} \tag{29}$$

which follow from (15). With $m_X = 2.41\times10^{-1}\,\text{V}^2\,\text{W}^{-1}$, $c_X = 4.35\times10^{-4}\,\text{V}^2$, $m_P = 2.33\times10^{-1}\,\text{V}^2\,\text{W}^{-1}$, $c_P = 5.60\times10^{-4}\,\text{V}^2$, and $P_{\text{LO}}^{(\text{det})} = 8.4\,\text{mW}$,

$$\Delta_{\text{dB}}^{(X)} \approx 7.52\,\text{dB}, \qquad \Delta_{\text{dB}}^{(P)} \approx 6.52\,\text{dB}. \tag{30}$$

One difference compared to the other references is the dip in intensity we see below 0.2 GHz, which is caused by the HPF we have added which has a bandwidth of 185 to 3000 MHz. The resulting filter band (185–1650 MHz) lies only slightly outside of Nyquist frequency for $f_{\text{s}} = 3.2\,\text{GHz}$ ($f_{\text{s}}/2 = 1.6\,\text{GHz}$), so residual aliasing into the analysis region is negligible. Another inconsistency is the observation of spikes in the spectrum. The spikes at the highest frequency are related to spurious codes of the ADC at $f_{\text{s}}/2$. The spikes at lower frequencies are related to electromagnetic interference resulting from surrounding electronics because they can be seen with LO off or on, indicating that their source is not related to the optics. These spikes are narrow in bandwidth and do not impact the security after extraction.

As a consistency check, we integrate the PSD difference (LO on minus LO off) over the analysis band (185–1600 MHz) to obtain the quantum-to-classical noise ratio (QCNR), reported with uncertainty. This cross-check is consistent with the slope-based calibration used for $\delta'_{X/P}$ and is not used directly in the entropy bound.

PSD estimates were obtained using a one-sided periodogram with a Hann window (equivalent noise bandwidth, ENBW $= 1.5\,\Delta f$). The fast Fourier transform length was set to $N = 2^{20}$, with a sampling rate of $f_{\text{s}} = 3.2\,\text{GS}\,\text{s}^{-1}$, resulting in a frequency bin width of $\Delta f = f_{\text{s}}/N$. Each spectrum was averaged over $M = 64$ blocks, and the error is $\pm 1.96/\sqrt{M}$ confidence interval on the dB scale, which is not shown due to being negligibly small.

To further characterize the QRNG, we measure the Husimi Q function and the variance dependence on LO power. With the vacuum-unit axes from equations (15)–(22), the 2D histogram H_{mn} over bin area $\delta'_X\delta'_P$ is converted via

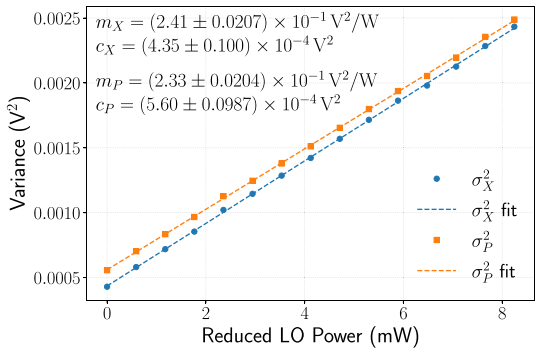

Figure 5. The dependence of both quadrature measured signal variance on the local oscillator laser power.

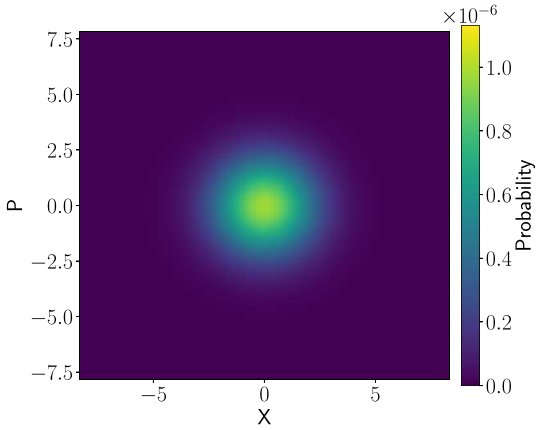

Figure 6. Husimi–Q distribution in quadrature space: horizontal axis shows the in-phase quadrature X and vertical axis the orthogonal quadrature P. The color scale encodes the normalized probability per pixel.

$$\widehat{Q}_{mn} := \frac{2H_{mn}}{N\delta'_X\delta'_P}, \qquad \sum_{m,n} \widehat{Q}_{mn}\,\frac{\delta'_X\delta'_P}{2} = 1. \tag{31}$$

The calibration curve is shown in figure 5. The resulting fit values are:

$$\begin{aligned} m_X &= (2.41 \pm 0.0207) \times 10^{-1}\,\mathrm{V^2/W}\ (95\%\,\mathrm{CI}),\\ c_X &= (4.35 \pm 0.100) \times 10^{-4}\,\mathrm{V^2}\ (95\%\,\mathrm{CI}),\\ m_P &= (2.33 \pm 0.0204) \times 10^{-1}\,\mathrm{V^2/W}\ (95\%\,\mathrm{CI}),\\ c_P &= (5.60 \pm 0.0987) \times 10^{-4}\,\mathrm{V^2}\ (95\%\,\mathrm{CI}). \end{aligned}$$

At the operating LO power $P_{\mathrm{LO}}^{(\mathrm{det})} = 8.4\,\mathrm{mW}$ we obtain dimensionless vacuum-unit variances (e.g. $\sigma_X^2 \simeq 0.616$, $\sigma_P^2 \simeq 0.647$) and the conservative resolutions

$$\delta'_X = 0.0300 \quad \text{and} \quad \delta'_P = 0.0319 \tag{32}$$

(with uncertainties), computed via equations (17)–(22). These resolutions are then used to convert the voltage histogram to the phase-space representation shown in figure 6. From these resolutions,

$$\begin{aligned} h_{\min}^{(1)} &= -\log_2\left(\min\left\{\tfrac{\delta'_X\delta'_P}{2\pi}, 1\right\}\right)\\ &= \log_2\left(\frac{2\pi}{0.0300 \times 0.0319}\right) \approx 12.681\ \text{bits/round}. \end{aligned} \tag{33}$$

For $n_{\mathrm{eff}} = k_{\mathrm{in}}/b = 2880/24 = 120$ rounds per block, this gives

$$H_{\min}\left(Z^{n_{\mathrm{eff}}}|E\right) \geqslant 1521\ \text{bits/block}. \tag{34}$$

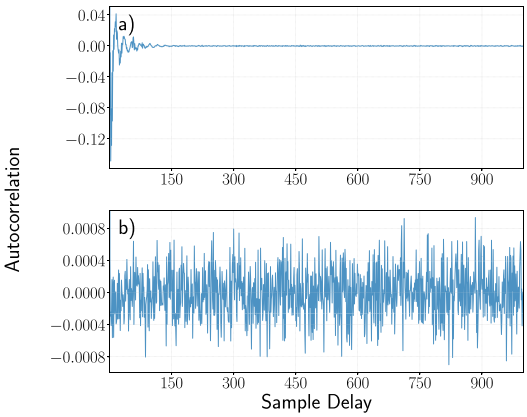

Figure 7. Autocorrelation comparison between (a) raw quadrature data and (b) extracted random bits. Panel (a) shows significant positive correlation at short lags, decaying over ≈3 samples consistent with the band-limited detection bandwidth ($\tau_c \sim 1/\Delta f$). Panel (b) shows residual correlations after Toeplitz extraction at the 10^{-4} level, indistinguishable from statistical noise, confirming successful removal of correlated structure. This comparison serves as a diagnostic verifying that the extractor operates as designed. Lag axis is in units of 10^3 samples; optical power 8.4 mW; 10^7 samples used.

We verified that the empirical bin frequencies f_{mn} satisfy $f_{mn} \leqslant \min\{\delta'_X\delta'_P/(2\pi), 1\} + \Delta$ with high confidence, using a multiplicative Chernoff bound at significance $\alpha = 10^{-6}$ together with a union bound over all bins (m, n); violations signal a calibration error in $(\alpha_{X/P}, \delta'_{X/P})$.

Several discrepancies can be seen compared to [10]. Firstly, we show an increased variance value in vacuum units. The most likely reason is the difference in set-ups. [10] immediately measures the quadrature signals from the diode with an oscilloscope that has a low enough measurement range to capture the low diode voltages. In our case, the lowest voltage that can be measured with the ADC without measurement quality deterioration is 0.5 V peak-to-peak (from −0.25 to 0.25 V), therefore, we have used an amplifier, which may have different noise characteristics, potentially causing an increase in the noise.

The second difference is that in our case the power dependencies of both quadratures do not completely overlap. This is also related to the difference in set-ups. Together with amplifiers we used hardware filters to reduce the complexity of the FPGA implementation. While the components models were the same between the two quadratures, there can still be variations between the same devices in insertion loss, gain, and other characteristics that can cause a reduction in the signal. To increase the overlap, we used the EVOAs before the photodiodes to reduce the power of one of the quadratures.

With the power calibration complete, we obtain $H_{\min} \geqslant 1521$ per block; we extract $\ell = 1272$ bits per block, i.e. $(\ell/k_{\text{in}}) \approx 0.442$, using a 1272×2880 Toeplitz matrix. The next step is to confirm whether the Toeplitz extraction algorithm is flattening the distribution seen in figure 6. This can be partially done by looking at the autocorrelation of the numbers before and after extraction, which is shown in figure 7. The gap between the min-entropy bound (1521 bits) and extracted length (1272 bits) is deliberate. The leftover hash lemma requires a privacy amplification penalty of $2\log_2(1/\varepsilon_{\text{PA}}) = 128$ bits for $\varepsilon_{\text{PA}} = 2^{-64}$, yielding a theoretical maximum of 1393 bits. We extract $\ell = 1272 = 24 \times 53$ bits, which aligns with the 24-bit word boundaries required by the parallel FPGA pipeline. The remaining 121-bit margin provides robustness against calibration drift during continuous operation.

The net extracted bit rate is

$$R_{\text{net}} = \frac{\ell}{n_{\text{eff}}} f_{\text{s}}^{(\text{eff})}, \tag{35}$$

where n_{eff} is the number of retained rounds after enforcing factorization (by fixed-ratio decimation or an invertible whitener). For decimation by D, $n_{\text{eff}} = n/D$ and $f_{\text{s}}^{(\text{eff})} = f_{\text{s}}/D$; for an orthonormal (unit-Jacobian) transform $n_{\text{eff}} = n$ and $f_{\text{s}}^{(\text{eff})} = f_{\text{s}}$. Clipped samples are dropped, so n_{eff} already counts valid rounds and no separate Δ_{clip} term is needed. The measured clipping fraction at the operating point is $f_{\text{clip}} = 2.3 \times 10^{-6}$ (95% CI).

There is a clear correlation observed in the random data before extraction, which is also seen in other QRNGs [9, 10, 31]. It is caused partially by noise introduced by the sampling device [10], however, the most significant contributor in this case is the sample rate related to the filtering that we apply. Band-limitation induces temporal correlations. We estimate the normalized ACF $\rho(k)$ and define the integrated autocorrelation time

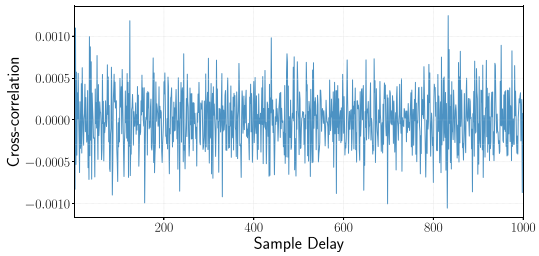

Figure 8. Cross-correlation between X and P quadratures of the raw data. The autocorrelation and lag is applied in terms of 1000 12-bit samples. The optical power is 8.4 mW. A total of 10^7 samples were taken for calculation.

$$\tau_{\text{int}} := \max\left\{1,\ 1+2\sum_{k=1}^{K^\star}\rho(k)\right\},$$
$$K^\star = \min\left\{k \geqslant 1 :\ |\rho(k)| < 2/\sqrt{N_{\text{samp}}}\right\}. \tag{36}$$

In our data we obtain

$$\tau_{\text{int}} = 2.68 \pm 0.00011\ (95\%\ \text{CI}),$$

using $B = 10^4$ bootstrap resamples with block length $\ell_B = 10^5$ to keep the calculation time reasonable. Nyquist-rate sampling ($f_s = 2\Delta f$) prevents aliasing but does not yield independent samples; approximate independence would require $f_s \lesssim \Delta f$, reducing throughput by at least half. We instead sample at $f_s = 3.2\,\text{GS}\,\text{s}^{-1}$ and let the extractor compression ratio $\ell/k_{\text{in}} \approx 0.44$ absorb the correlation structure. This is consistent with the security model: the per-round bound derives from the POVM operator norm (lemma 1), which holds for any input state including states correlated with previous rounds. The accumulation $H_{\min}(Z^n|E) \geqslant n \cdot h_{\min}^{(1)}$ requires only that the trusted measurement applies identical non-adaptive POVMs, a condition satisfied by fixed ADC discretization regardless of temporal correlations. The quantity τ_{int} thus serves as a diagnostic confirming the analog chain behaves as characterized, not as an input to the min-entropy bound. After extraction, residual autocorrelations are consistent with statistical fluctuations (we do not reduce the security bound by τ_{int} since it can be minimized by decimating the sampling frequency). We inspected higher-order autocorrelations (Ljung–Box) up to a fixed lag window, finding no statistically significant residual structure at the 95% level. The experimental cross-quadrature correlation was $|\rho_{XP}| = 0.012 \pm 0.004$, consistent with the negligible correlation penalty reported in section II. This completes the security-relevant calibration; the min-entropy bound and extraction parameters are now fully determined.

3.2. Statistical diagnostics

The following tests serve as implementation diagnostics to verify that the hardware operates as modeled. We emphasize that statistical tests are necessary but not sufficient for security: a deterministic pseudorandom number generator with sufficient state size will pass all such tests while possessing zero min-entropy against an adversary who knows the algorithm. The security of our QRNG is established by the calibration and min-entropy analysis of section 3.1, not by the tests below.

We additionally inspected cross-correlation between X and P quadratures of the raw data to confirm there is no correlation introduced by the ADC sampling or the electrical filtering. The correlation computation was performed in the same method as for autocorrelation. The result can be seen in figure 8, which shows that the cross-correlation between X and P is negligible.

Statistical quality of the generated random numbers was evaluated using both the GNU Dieharder suite [32] and the NIST SP 800–22 Statistical Test Suite [33]. For Dieharder, all available tests were executed on binary input files of size 16 GB, with the generator set to file input and the analysis configured to apply the Kolmogorov–Smirnov test with two rejections for failure determination and full reporting of p-values. We also sorted the resulting p-values from the Dieharder tests and plotted them on a Q–Q plot. For the NIST tests, we run it configured to do 1000 iterations with 100 p-values per iteration, using sequences of 1000 000 bits each. The results from the tests are shown in figure 9. As can be seen, the NIST tests are all passing and the dieharder test p-values are all uniformly distributed along the diagonal reference line, indicating that no patterns were detected. These tests confirm that no obvious implementation defects (stuck bits, periodic patterns, unexpected correlations from the analog chain) are

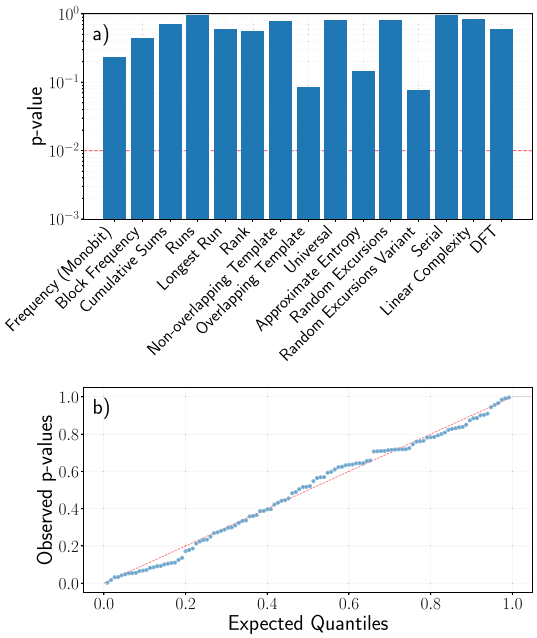

Figure 9. Statistical test results for quantum random number generator output. (a) NIST Statistical Test Suite results showing p-values for individual randomness tests on a logarithmic scale. The dashed red line indicates the significance threshold ($\alpha = 0.01$). Tests with single occurrences show individual p-values, while tests with multiple instances show Kolmogorov–Smirnov test p-values comparing the distribution of individual p-values against a uniform distribution. (b) Quantile–quantile plot of Dieharder test suite p-values against expected uniform quantiles. Points following the diagonal reference line (dashed red) indicate p-values consistent with a uniform distribution expected from a true random source.

present. However, passing statistical tests does not establish security—the information-theoretic guarantees derive solely from the SDI framework and calibration analysis of section.

4. Conclusions

We have presented a fully hardware-integrated, SDI QRNG based on heterodyne detection of vacuum fluctuations. The SDI framework, which derives security from the POVM structure rather than source characterization, was established by [10]; our contribution lies in extending the security analysis to address real-time operational requirements (clipping, excess noise, temporal correlations, calibration uncertainty) and demonstrating the first online SDI-certified extraction at rates exceeding 30 Gb s^{-1}. We realize the random number acquisition and Toeplitz hashing based randomness extraction entirely on FPGA logic. We also do an extensive security analysis, which shows that the source-device independent model that we consider does not rely on a model of the source, but rather on the properties of the measurement itself, which is useful in the case where the source characteristics drift or when they may be adversarially controlled.

We demonstrate a system with a real-time random number generation rate of 33.92 Gbit s^{-1} with all the security bounds taken into account. To quantify this claim, we decompose the contributions to the certified rate. The single-round min-entropy $h_{\min}^{(1)} = \log_2(2\pi/\delta_X'\delta_P') \approx 12.68$ bits/round receives the following corrections: (i) excess noise inflation reduces this by $\Delta h_\gamma = \log_2(\gamma_X\gamma_P) \approx 0.34$ bits/round (-2.7%); (ii) clipping contributes $< 10^{-5}$ bits/round (negligible); (iii) cross-correlation penalty is $< 10^{-4}$ bits/round (negligible). In contrast, the raw bin width scales as $\delta \propto 2^{-n_{\text{ENOB}}}$, so a 1-bit improvement in ENOB would increase $h_{\min}^{(1)}$ by 2 bits/round ($+15.8\%$). The ENOB contribution thus dominates other effects by a factor of ≈ 6, confirming that higher-ENOB ADCs offer the most direct path to increased rates. We additionally explore additional security limiting factors, such as ADC clipping, inter-quadrature correlation and time autocorrelation. We find that these quantities do not significantly impact the security and therefore we only present them as diagnostics.

Most importantly, this set-up shows that high bandwidth true RNG can be achieved with commonly available components. Moreover, the PCIe interface allows for high speed transfer of the random numbers into another platform, allowing easy access to the QRNG for different applications.

Acknowledgments

The authors acknowledge the Bundesministerium für Bildung und Forschung in the frame of the project QR.N (contract no. 16KIS2201) and EPSRC Grant No. EP/S030751/1. A. G. acknowledges funding by the state of North Rhine-Westphalia through the EIN Quantum NRW program. Data is made available upon request.

Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

Author contributions

Marius Cizauskas 0000-0003-1567-2048
Conceptualization (equal), Data curation (lead), Formal analysis (supporting), Investigation (lead), Methodology (lead), Software (lead), Validation (lead), Visualization (lead), Writing – original draft (lead), Writing – review & editing (equal)

Hamid Tebyanian 0000-0002-9887-4130
Conceptualization (equal), Formal analysis (lead), Investigation (equal), Methodology (equal), Writing – original draft (equal), Writing – review & editing (equal)

A Mark Fox 0000-0002-9025-2441
Supervision (supporting), Writing – review & editing (equal)

Manfred Bayer 0000-0002-0893-5949
Project administration (supporting), Writing – review & editing (equal)

Marc Assmann 0000-0003-4953-1286
Conceptualization (supporting), Investigation (supporting), Project administration (supporting), Writing – review & editing (equal)

Alex Greilich 0000-0001-7813-682X
Conceptualization (lead), Funding acquisition (lead), Methodology (equal), Project administration (lead), Resources (lead), Supervision (lead), Writing – original draft (supporting), Writing – review & editing (equal)

References

[1] Martín V *et al* 2021 Quantum technologies in the telecommunications industry *EPJ Quantum Technol.* **8** 19
[2] Alkhazragi O, Lu H, Yan W, Almaymoni N, Park T-Y, Wang Y, Ng T K and Ooi B S 2023 Semiconductor emitters in entropy sources for quantum random number generation *Ann. Phys., Lpz.* **535** 2300289
[3] Başar E 2023 Kirchhoff meets johnson: in pursuit of unconditionally secure communication *Eng. Rep.* **6** e12958
[4] Garipcan A M and Erdem E 2021 Hardware implementation of chaotic zigzag map based bitwise dynamical pseudo random number generator on field-programmable gate array *Informacije MIDEM - Journal of Microelectronics, Electronic Components and Materials*
[5] Jennewein T, Achleitner U, Weihs G and Weinfurter H (of Vienna A Z U, of Innsbruck U and of Munich U) 1999 A fast and compact quantum random number generator *Rev. Sci. Instrum.* **71** 1675
[6] Yang J, Liu J, Su Q, Li Z, Fan F, Xu B and Guo H 2016 5.4 gbps real time quantum random number generator with simple implementation *Opt. Express* **24** **24** 27475
[7] Williams C R S, Salevan J, Li X, Roy R and Murphy T E 2010 Fast physical random number generator using amplified spontaneous emission *Opt. Express* **18** **23** 23584
[8] Bai B, Huang J, Qiao G-R, Nie Y-Q, Tang W, Chu T, Zhang J and Pan J-W 2021 18.8 Gbps real-time quantum random number generator with a photonic integrated chip *Appl. Phys. Lett.* **118** 264001
[9] Zheng Z, Zhang Y, Huang W, Yu S and Guo H 2019 6 gbps real-time optical quantum random number generator based on vacuum fluctuation *Rev. Sci. Instrum.* **90** 043105
[10] Avesani M, Marangon D G, Vallone G and Villoresi P 2018 Source-device-independent heterodyne-based quantum random number generator at 17 Gbps *Nat. Commun.* **9** 5365
[11] Bruynsteen C, Gehring T, Lupo C, Bauwelinck J and Yin X 2023 100-Gbit/s integrated quantum random number generator based on vacuum fluctuations *PRX Quantum* **4** 010330
[12] Bertapelle T *et al* 2025 High-speed source-device-independent quantum random number generator on a chip *Opt. Quantum* **3** 111
[13] Qiu K, Cai Y, Ng N H Y and Haw J Y 2025 Fully passive quantum random number generation with untrusted light *APL Quantum* **2** 046105

[14] Cheng J, Liang S, Qin J, Li J, Yan Z, Jia X, Xie C and Peng K 2024 Semi-device-independent quantum random number generator with a broadband squeezed state of light *npj Quantum Inf.* **10** 20
[15] Zhang J, Li Y, Zhao M, Han D, Liu J, Wang M, Gong Q, Xiang Y, He Q and Su X 2025 One-sided device-independent random number generation through fiber channels *Light Sci. Appl.* **14** 25
[16] Yang J *et al* 2025 An ultra-fast quantum random number generation scheme based on laser phase noise *Commun. Phys.* **9** 5
[17] Crampton O M *et al* 2025 A 2-gbps low-SWaP quantum random number generator with photonic integrated circuits for satellite applications *npj Quantum Inf.* **11** 153
[18] Grünenfelder F, Boaron A, Rusca D, Martin A and Zbinden H 2020 Performance and security of 5 GHz repetition rate polarization-based quantum key distribution *Appl. Phys. Lett.* **117** 144003
[19] Takesue H, Nam S W, Zhang Q, Hadfield R H, Honjo T, Tamaki K and Yamamoto Y 2007 Quantum key distribution over a 40-db channel loss using superconducting single-photon detectors *Nat. Photon.* **1** 343
[20] Ng S Q, Zhang G, Wang C and Lim C C W 2023 240 Gbps quantum random number generator with photonic integrated chip *Conf. on Lasers and Electro- Optics (CLEO)*
[21] Thewes J, Lüders C and Amann M 2019 Eavesdropping attack on a trusted continuous-variable quantum random-number generator *Phys. Rev. A* **100** 052318
[22] Liu W-Z *et al* 2019 Device-independent randomness expansion against quantum side information *Nat. Phys.* **17** 448
[23] Shrivastava M, Mittal M, Kumari I and Abhignan V 2025 Randomness in quantum random number generator from vacuum fluctuations with source-device-independence *Appl. Phys. B* **131** 111
[24] Guo X, Zhou W, Luo Y, Meng X, Li J, Bian Y, Guo Y and Xiao L 2025 Deep learning-based min-entropy-accelerated evaluation for high-speed quantum random number generation *Entropy* **27** 786
[25] Li Y, Fei Y, Wang W, Meng X, Wang H, Duan Q, Han Y and Ma Z 2023 Practical security analysis of a continuous-variable source-independent quantum random number generator based on heterodyne detection *Opt. Express* **31** 23813
[26] Nie Y-Q, Huang L, Liu Y, Payne F, Zhang J and Pan J-W 2015 The generation of 68 Gbps quantum random number by measuring laser phase fluctuations *Rev. Sci. Instrum.* **86** 063105
[27] Tomamichel M, Schaffner C, Smith A and Renner R 2011 Leftover hashing against quantum side information *IEEE Trans. Inf. Theory* **57** 5524
[28] Krawczyk H 1994 Lfsr-based hashing and authentication *Advances in Cryptology — Crypto'94* ed Y G Desmedt (Springer) pp 129–39
[29] AMD 2025 UltraFast design methodology guide (UG949)
[30] Shen Y, Tian L and Zou H 2010 Practical quantum random number generator based on measuring the shot noise of vacuum states *Phys. Rev. A* **81** 063814
[31] Li J, Huang Z, Yu C, Wu J, Zhao T, Zhu X and Sun S 2024 Quantum random number generation based on phase reconstruction *Opt. Express* **32 4** 5056
[32] Brown R G 2006 *Dieharder: A GNU Public License Random Number Tester* (*Physics Department*) (Duke University)
[33] Bassham L E *et al* 2010 A statistical test suite for random and pseudorandom number generators for cryptographic applications *Technical Report* Special Publication 800-22 Rev. 1a National Institute of Standards and Technology