On-chip semi-device-independent quantum random number generator exploiting contextuality: Supplementary Information

Maddalena Genzini,[1] Caterina Vigliar,[1] Mujtaba Zahidy,[1] Hamid Tebyanian,[2] Andrzej Gajda,[3] Klaus Petermann,[4] Lars Zimmermann,[3,5] Davide Bacco,[6,7] and Francesco Da Ros[1]

[1] *Department of Electrical and Photonics Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark*
[2] *School of Physical and Chemical Sciences, Queen Mary University of London, London, United Kingdom*
[3] *IHP – Leibniz-Institut für innovative Mikroelektronik, Frankfurt (Oder), Germany*
[4] *Technische Universität Berlin, Institut für Hochfrequenz- und Halbleiter-Systemtechnologien, Berlin, Germany*
[5] *Technische Universität Berlin, FG Silizium-Photonik, Berlin, Germany*
[6] *University of Florence, Department of Physics and Astronomy, Sesto Fiorentino, Italy*
[7] *QTI s.r.l., Firenze, Italy*

SUPPLEMENTARY SECTION I: EXPERIMENTAL SETUP

The experimental setup for the source is shown in Fig. 1. A pump source generates a pair of photons via spontaneous four wave mixing (SFWM) in a silicon waveguide. Here the pump photons are converted into two (signal and idler) photons at different frequencies, at symmetric positions in the spectrum with respect to the pump photons.

In order to make a comprehensive photon-pair generation study and achieve the optimal generation rates for our

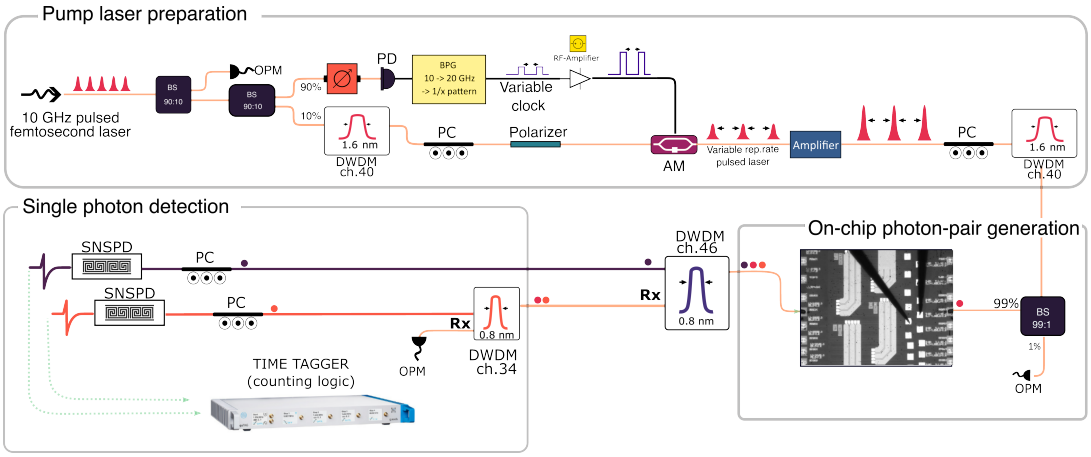

FIG. 1: Schematics of the set-up for photon-pair source characterization. The system can be divided into three stages: laser pump preparation, on-chip photon-pair generation and single photon detection.

silicon structures, an initial pulsed laser (5-ps pulses) at 1545.32 nm with a fixed ∼10 GHz repetition rate was pulse-picked to ∼1.25 GHz and the pair-generation rate has been investigated. The setup to tune the repetition rate of the laser is shown in Fig. 1. The laser is split with a 90%10% splitter: the 90% output is used to generate a synchronized variable clock that can be used to modulate the 10% output in order to achieve variable repetition rates. In more detail, the 90% output is detected by a 50 GHz photodetector providing a clock for a bit pattern generator (BPG). The BPG allows defining a custom bit pattern, syncronized to the rate of the pulsed laser, which can then be amplified and sent to a Mach-Zehnder modulator (MZM). By adjusting the bit pattern in the BPG we can suppress pulses in the pulse train from the original lasers, therefore tuning its repetition rate. The 10% output is preliminary filtered by a single channel DWDM filter with 1.6 nm width and then connected to the optical input of the MZM. In order to optimize its performance a polarization controller and polarizer are inserted at its input as well. The pulse train at the modulator's output is then amplified by an erbium doped fiber amplifier. The spectral background was removed through the use of a second single channel DWDM filter with 1.6 nm bandwidth. Just before the chip's input port, an off-chip pump power of ≈ 6 dBm was used in typical measurement conditions, monitored by the 1% port of an auxiliary 99%1% splitter. A fiber polarization controller (PC) is also used before the silicon device to maximize the coupling into the chip. Once coupled into the chip, the laser light is used to pump a 2 cm long waveguide. The chip used for the photon-pair generation was fabricated in a SiGe:BiCMOS line at IHP - Leibniz Institute for High Performance Microelectronics [1]. We found the optimal condition for the photon-pair generation pumping a 2 cm long waveguide with pulses at a repetition rate of 1.25 GHz. To enhance FWM efficiency free-carrier absorption is suppressed through a reverse biased P-I-N junction [1, 2], reaching an improvement in the pair generation rate, shown in Fig. 2 left. The pumped source generates a pair of photons: the wavelengths of the signal and idler photons are selected at 1541.35 nm (ITU channel 45) and 1550.12 nm (ITU channel 34), respectively. The output photons are coupled into single-mode fiber through the output grating couplers. Off-chip filtering of the spurious pump light was achieved via two DWDM filters (one per signal/idler branch) with a ≈ 0.8 nm bandwidth and > 50 dB extinction. In order to characterize the quantum source, photons were finally detected using two superconducting nanowire single photon detectors (SNSPDs), with system detection efficiency of 80%, < 100 Hz dark counts, approximately 70 ps FWHM timing jitter and recovery times of approximately ∼ 40 ns. Fast counting logic (timetagger), supporting more than 20 million events per second, was used to collect the single-photon detections and process them on the fly.

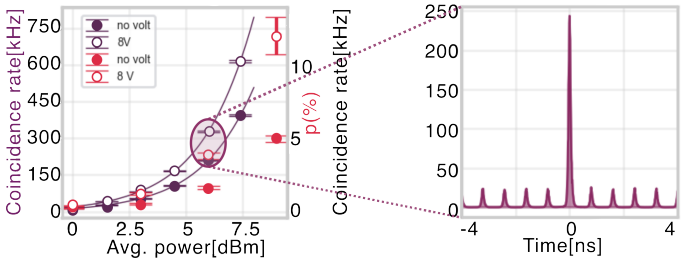

FIG. 2: **Left)** Coincidental counts vs avg power and probability of generation. **Right)** Histogram of chosen operational condition.

Heralded photon source

For best operation in our protocol, we select a photon-pair generation probability of $p \sim 3\%$, by keeping the pump power around 6 dBm (highlighted in Fig. 2 left). This low probability minimizes the occurrence of multi-pair events, which is critical in SFWM-based sources to ensure the generation of predominantly single photon pairs (Fig. 2 right) [3].

SUPPLEMENTARY SECTION II: QUANTUM CONTEXTUALITY

The Kochen–Specker theorem is a fundamental "no-go" theorem in the foundations of quantum mechanics. It shows that for quantum systems of dimension three or higher, it is impossible to assign definite, pre-existing values to all possible measurements (observables) in a non-contextual way, i.e., in a way that does not depend on which order measurements are being performed simultaneously [4, 5]. In a classical world, measurement outcomes are assumed to be predetermined and independent of other measurements, which is, mathematically speaking, equivalent to the existence of a joint probability distribution describing all observable properties. Quantum mechanics, however, rules out such a description: the outcome of a measurement cannot be interpreted as the unveiling of a pre-existent value, but rather as the result of an interaction between the measuring device and the system. This fundamental concept is known as contextuality, and it can be revealed experimentally through violations of non-contextual hidden variable (NCHV) inequalities [6]. In this work, we focus on the Klyachko–Can–Binicioğlu–Shumovsky (KCBS) inequality [7], which involves the use of the simplest state in which the contradiction is possible, a qutrit.

KCBS inequality

The KCBS inequality [7], also known as the pentagram inequality, is a fundamental inequality in the study of quantum contextuality for systems with a three-dimensional quantum state (qutrits). We consider five dichotomic observables A_1, A_2, A_3, A_4, A_5, each with outcomes ± 1. According to NCHV theory, if these observables admit a joint probability distribution, the following inequality holds:

$$\langle \chi_{\mathrm{KCBS}} \rangle = \langle A_1 A_2 \rangle + \langle A_2 A_3 \rangle + \langle A_3 A_4 \rangle + \langle A_4 A_5 \rangle + \langle A_5 A_1 \rangle \geq -3. \tag{1}$$

The inequality in (1) is known as the KCBS inequality, and provides a way to test for contextuality by setting a limit on the correlations that can be observed in case the system were non-contextual. Violation of this bound implies the impossibility of a joint probability distribution for the outcomes, and thus demonstrates contextuality.

Following the method explained in [8], the inequality in Eq. (1) could be verified under the assumption that conducting an extensive series of individual experiments leads to a fair sampling of a joint probability distribution and that the measurement outcomes of $< A_i A_j >$ are not influenced by the order in which the measurements are carried out.

The corresponding measurement scheme is depicted in Fig. 3. It consists of an initial state preparation stage, where a single photon is distributed across three optical modes (Fig. 3.i). The state preparation is followed by one of the five measurement stages implementing transformations T_i (Fig. 3.ii-vi). Each T_i mixes only on two of the modes, leaving the third unaffected, ensuring pairwise compatibility of the corresponding observables.

To implement the different stages, we used a reconfigurable hexagonal interferometer mesh, which implements arbitrary unitary transformations in the qutrit space.

Observables A_i

The five observables entering the KCBS inequality correspond to projective measurements onto five qutrit states $\{|\psi_i\rangle\}_{i=1}^{5}$, embedded in a three-dimensional Hilbert space spanned by the optical path basis $\{|0\rangle, |1\rangle, |2\rangle\}$. Each observable is defined as

$$A_i = \mathbb{I} - 2\,|\psi_i\rangle\langle\psi_i|, \tag{2}$$

so that the outcome -1 corresponds to projection onto $|\psi_i\rangle$, while $+1$ corresponds to projection onto its orthogonal subspace. The KCBS compatibility condition,

$$\langle \psi_i | \psi_{i+1} \rangle = 0 \qquad (i = 1, \ldots, 5 \bmod 5), \tag{3}$$

ensures that neighboring observables can be jointly measured.

In our implementation, the input qutrit state is a coherent superposition of the three optical paths,

$$|\psi_{\mathrm{in}}\rangle = \frac{1}{\sqrt[4]{5}} \left(|0\rangle + |1\rangle \right) + \sqrt{1 - \frac{2}{\sqrt{5}}}\, |2\rangle, \tag{4}$$

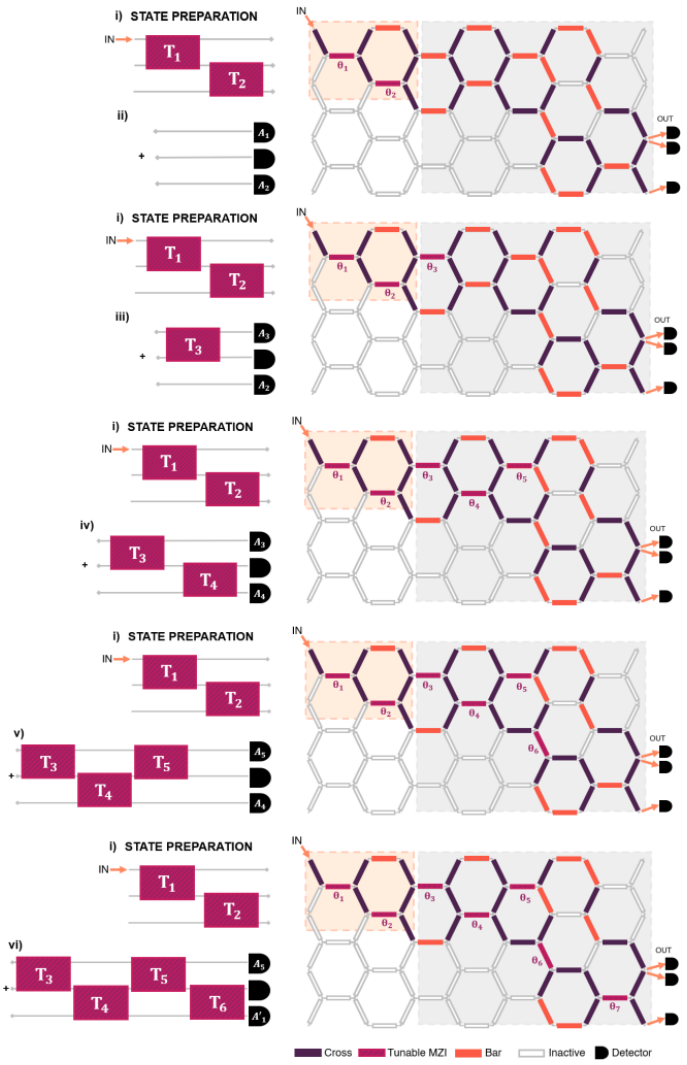

FIG. 3: Measurements scheme and on-chip implementation for all the different contexts.

which can be viewed as a rotated version of the ideal KCBS state $|2\rangle$ in the canonical theoretical frame. Accordingly, the measurement states $|\psi_i\rangle$ are obtained by applying the same unitary rotation U to the canonical KCBS vectors $|v_i\rangle$:

$$|\psi_i\rangle = U\,|v_i\rangle, \qquad A_i = \mathbb{I} - 2\,|\psi_i\rangle\langle\psi_i|. \tag{5}$$

Experimentally, the states $|\psi_i\rangle$ are realized as superpositions of optical paths, with phases implemented via phase shifters inside the interferometer mesh. The rotations between different measurement bases are achieved by programming the mesh to apply the required transformations T_i, each acting non-trivially only on two modes. The response of two detectors monitoring the optical modes defines a pair of measurement outcomes as appearing in the KCBS inequality (Eq. 1). The outcomes of the measurements (A_i) are determined by asking whether the detector has registered a click or not. If a detector clicks (or does not), then a value of -1 $(+1)$ is assigned to the corresponding measurement. As a result, the KCBS measurement structure has a natural implementation in a three-mode photonic circuit, with the outcomes determined by single-photon detections at the output modes. Randomness can be extracted from the outcomes of these projective measurements: data is used to estimate $\langle\chi_{\mathrm{KCBS}}\rangle$ and verify contextuality, while the remainder forms the raw randomness that is subsequently processed into uniform random bits.

SUPPLEMENTARY SECTION III: QUANTUM STATE TOMOGRAPHY

Quantum state tomography reconstructs an unknown quantum state by performing a set of measurements on multiple identically prepared copies of the state. In the present work, a qutrit state $|\Psi\rangle$ is prepared, and the corresponding density matrix ρ is reconstructed to verify consistency with the target state. A qutrit occupies a Hilbert space of dimension $d = 3$, hence ρ is a 3×3 positive semidefinite matrix with unit trace.

In this work, the *maximum likelihood estimation* (MLE) method has been used, as shown in previous works on tomography [9]. The necessary steps for this method can be summarized as follows:

1. The state is written in a physical density matrix representation. Any state can be written in this representation if the basis used is Hermitian, positive, and normalized. This density matrix will be a function of several unknown parameters.
2. Several mutually unbiased (MU) bases are measured, allowing us to define a likelihood function quantifying how well the parameterized density matrix matches the experimental data.
3. The likelihood function is then optimized to find the parameters used to rewrite the state in that representation.

Gell-Mann representation

For the first step, we write the state in the Gell-Mann representation. The Gell-Mann matrices $\{\lambda_i\}$ are a set of eight linearly independent 3×3 traceless Hermitian matrices. Thus, any state of dimension $d = 3$ can be expanded in this basis. Hence, we write:

$$\rho = \frac{I_3}{3} + \frac{1}{2}\sum_{i=1}^{8} r_i \lambda_i, \tag{6}$$

where I_3 is the identity matrix, c_i are the real parameters we want to estimate via the likelihood function, and λ_i are the Gell-Mann matrices.

Measurement bases

Three mutually unbiased (MU) bases are used to extract the information required for the reconstruction. Two orthonormal bases $\{|a_i\rangle\}$ and $\{|b_j\rangle\}$ in dimension d are MU if $|\langle a_i|b_j\rangle|^2 = 1/d$ for all i, j. Measurements in MU bases are informationally complete for qutrit tomography [10].

The three bases employed are:

- the computational basis $\{|0\rangle, |1\rangle, |2\rangle\}$,
- the Fourier basis,
- a third MU basis constructed from the Fourier basis.

Computational basis. The computational basis is defined as

$$|0\rangle = \begin{pmatrix}1\\0\\0\end{pmatrix},\ |1\rangle = \begin{pmatrix}0\\1\\0\end{pmatrix},\ |2\rangle = \begin{pmatrix}0\\0\\1\end{pmatrix},$$

with projectors $M_k = |k\rangle\langle k|$ for $k = 0, 1, 2$.

Fourier basis. The discrete Fourier transform (DFT) basis is given by

$$|f_k\rangle = \frac{1}{\sqrt{3}}\sum_{j=0}^{2} \omega^{jk} |j\rangle, \qquad \omega = e^{2\pi i/3},\ k = 0, 1, 2.$$

The corresponding projectors are $M_k = |f_k\rangle\langle f_k|$.

Third MU basis. A third basis that is MU with respect to both the computational and Fourier bases can be chosen as [10]

$$|v_k\rangle = \frac{1}{\sqrt{3}} \sum_{j=0}^{2} \omega^{(j+k)^2} |j\rangle , \qquad k = 0, 1, 2.$$

The associated projectors are $M_k = |v_k\rangle \langle v_k|$.

Tomography measurements and maximum likelihood estimation method

The set of measurements used to extract the density matrix consists of 9 measurement runs (3 for each basis), each lasting 10 min. Averages of coincidental counts are summarized in Table I.

	Computational	**Fourier**	**MU**
1st	66 ± 8	131 ± 11	126 ± 11
2nd	11 ± 3	4 ± 2	4 ± 2
3rd	66 ± 8	4 ± 2	4 ± 2

TABLE I: Data averages from the tomography measurements (coincidental counts).

The average number of coincidental counts that will be observed in a given experimental run is

$$n_k \;=\; \eta \, \langle v_k| \, \rho \, |v_k\rangle , \quad \text{(Eq. 1)} \tag{7}$$

where ρ is the density matrix describing the prepared state and η is a constant dependent on the efficiency of the detectors. The projection measurement is represented by the operator

$$M_k \;=\; |v_k\rangle \langle v_k| ,$$

which is different for each experimental run.

Computational Basis. For the computational basis,

$$M_{k+1} \;=\; |k\rangle \langle k| , \quad k = 0, 1, 2,$$

with projectors $|k\rangle \langle k|$.

Fourier Basis. For the Fourier basis,

$$M_{k+4} \;=\; |v_{F_k}\rangle \langle v_{F_k}| ,$$

where

$$|v_{F_k}\rangle \;=\; \frac{1}{\sqrt{3}} \sum_{j=0}^{2} e^{2\pi i \, j \, k/3} \, |j\rangle , \quad k = 0, 1, 2.$$

Third MU Basis. For the third basis,

$$M_{l+7} \;=\; |v_{MU_l}\rangle \langle v_{MU_l}| ,$$

where

$$|v_{MU_l}\rangle \;=\; \frac{1}{\sqrt{3}} \sum_{j=0}^{2} \omega^{(j+k)^2} |j\rangle , \quad \omega = e^{2\pi i/3}, \quad k = 0, 1, 2.$$

Thus, the probability of the set of 9 measured expectation values $\{\overline{n}_1, \ldots, \overline{n}_9\}$ is

$$P(n_1, \ldots, n_9) \;=\; \frac{1}{N_{\rm norm}} \prod_{k=1}^{9} \exp\Big[-\frac{(\,n_k \;-\; \overline{n}_k\,)^2}{2\,\sigma_k^2}\Big],$$

where σ_k is the standard deviation of the kth coincidence (assumed Gaussian, $\sigma_k = \sqrt{n_k}$) and $N_{\rm norm}$ is a normalization constant.

The likelihood that the matrix $\rho(c_1, \ldots, c_8)$ could produce the measured data $\{\overline{n}_1, \ldots, \overline{n}_9\}$ is then

$$L(c_1, \ldots, c_8) \;=\; \frac{1}{N_{\rm norm}} \prod_{k=1}^{9} \exp\left[-\frac{\big(N\;\langle v_k|\,\rho(c_1, \ldots, c_8)\,|v_k\rangle \;-\; \overline{n}_k\big)^2}{2\,N\;\langle v_k|\,\rho(c_1, \ldots, c_8)\,|v_k\rangle}\right],$$

where N_{norm} is a normalization constant corresponding to the total number of photons in each basis. Rather than finding the maximum of $L(c_1, \ldots, c_8)$ directly, it is simpler to minimize the negative of its logarithm. Hence the optimization problem reduces to finding the minimum of:

$$L(c_1, \ldots, c_8) \;=\; \frac{1}{N} \sum_{k=1}^{9} \Big[-\frac{\big(N\;\langle v_k|\,\rho(c_1, \ldots, c_8)\,|v_k\rangle \;-\; \overline{n}_k\big)^2}{2\,N\;\langle v_k|\,\rho(c_1, \ldots, c_8)\,|v_k\rangle}\Big].$$

Minimization yields the optimal values of the parameters $(c_1, \ldots, c_8)$ and thus the reconstructed density matrix. The theoretical target density matrix is

$$\rho_{\rm theory} = \begin{pmatrix} 0.447 & 0.217 & 0.447 \\ 0.217 & 0.106 & 0.217 \\ 0.447 & 0.217 & 0.447 \end{pmatrix}.$$

The MLE optimization returns the parameters

$$(c_1, \ldots, c_8) = [\,0.221,\; 0.000,\; 0.443,\; -0.000,\; 0.229,\; -0.000,\; 0.221,\; -0.155\,],$$

which correspond to the reconstructed density matrix

$$\rho_{\rm exp} = \begin{pmatrix} 0.426 & 0.221 & 0.443 \\ 0.221 & 0.114 & 0.229 \\ 0.443 & 0.229 & 0.460 \end{pmatrix} \qquad \sigma\rho_{\rm exp} = \pm \begin{pmatrix} 0.040 & 0.026 & 0.023 \\ 0.026 & 0.025 & 0.025 \\ 0.023 & 0.025 & 0.041 \end{pmatrix}.$$

$$\rho_{\rm theory} - \rho_{\rm exp} = \begin{pmatrix} 0.021 & -0.004 & 0.004 \\ -0.004 & -0.008 & -0.012 \\ 0.004 & -0.012 & -0.013 \end{pmatrix}.$$

The good agreement between the theoretical prediction and the experimentally reconstructed state confirms the reliability of our tomographic procedure and the robustness of the experimental implementation. In particular, all element-wise deviations between $\rho_{\rm theory}$ and $\rho_{\rm exp}$ lie well within the corresponding experimental uncertainties $\sigma\rho_{\rm exp}$, estimated via Monte Carlo bootstrapping of the measured data, demonstrating that the reconstructed state is statistically consistent with the theoretical values.

SUPPLEMENTARY SECTION IV: COMPATIBILITY LOOPHOLE

The modified KCBS inequality framework requires a quantitative estimate of the compatibility between the observables A_1 and A_1' that are associated with the same vertex of the contextuality pentagram. Ideally, these two observables should correspond to identical measurements performed in different contexts, implying perfect compatibility. However, in our photonic integrated circuit, A_1 and A_1' are realized on the same optical mode and differ only by the final measurement setting in two distinct stages of the protocol. As a result, they cannot be jointly measured within a single pulse, and the compatibility must instead be inferred from their action on the same quantum state in separate rounds.

Following Refs. [11, 12], we introduce the experimentally accessible compatibility parameter

$$R_{\text{comp}} = \langle A_1 A_1' \rangle_{\text{obs}},$$

which quantifies how closely the two observables behave as the same measurement. Bounding R_{comp} is essential for security: if this overlap were left unconstrained, an adversary could exploit deviations from compatibility to increase their guessing probability, causing the certified min-entropy to collapse to zero.

To estimate R_{comp}, we compare outcomes obtained when measuring A_1 in Stage 1 (context A_1A_2) and when measuring A_1' in Stage 5 (context A_5A_1'). Since the corresponding datasets were acquired independently, we first recover the experimental clock from the heralding channel (D1), determine the pump repetition period, and temporally align the two measurement sequences on a pulse-by-pulse basis. A schematic pipeline of this procedure is shown below:

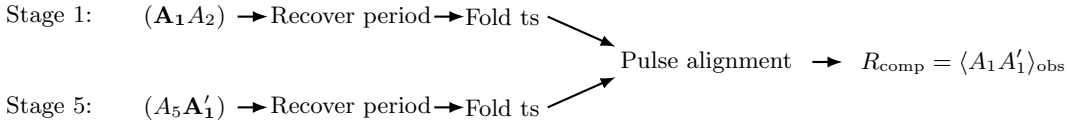

FIG. 4: **Processing pipeline for compatibility estimation.** Measurement outcomes for $\mathbf{A_1}$ (Stage 1) and $\mathbf{A_1'}$ (Stage 5) undergo clock-period recovery and timestamp folding, then are aligned pulse-by-pulse to evaluate $R_{\text{comp}} = \langle A_1 A_1' \rangle_{\text{obs}}$.

Clock recovery

The raw data consist of timestamps of photon detection events registered by four inputs corresponding to the four single-photon detectors (D1-D4). We performed a clock recovery technique based on [13], and temporally aligned to detection events for further processing.

The clock period estimation was done over a window of 500 μs and looped over the data acquired by detector D1 assigned to the heralding photons due to higher flux of photons. This clock period would also be valid for the other three channels as well. The clock recovery yields a stable value of approximately 803.75 ps. This recovered period defines the fundamental clock of the experiment.

Each detection timestamp t was then folded into one clock cycle using

$$t_{\text{folded}} = t \bmod T,$$

where T is the recovered period. For visualization and alignment, we chose a reference of 400 ps within the cycle as the center of the detection window. A systematic clock offset and drift across data acquisition is expected in this method as a result of imprecise Fast-Fourier transform, which is then corrected.

Coincidence counting within and across stages

Once both datasets were aligned to the same clock and delay-corrected, we performed coincidence counting. Within each stage, we counted detection events that occurred within the same clock window between the herald channel (D1) and each of the measurement channels (D2–D4). Each detection timestamp t is folded modulo the recovered period $T = 803.75$ ps, and a constant offset is corrected to match the detection windows across the two stages. Across

stages, we identified common pulse windows by requiring that detections occurred within a fraction (e.g. $T/10$) of the recovered clock period. For each aligned pump pulse k, a dichotomic outcome $a_1^{(k)} \in \{+1, -1\}$ is assigned in Stage 1 depending on which detector fired, and similarly $a_1'^{(k)}$ in Stage 5. For each pair of aligned pulses, we then assigned outcome values a_1 and a_1' depending on which detector fired, and computed their product. The compatibility is then obtained as

$$R_{\text{comp}} = \frac{1}{N} \sum_{k=1}^{N} a_1^{(k)} a_1'^{(k)},$$

with binomial uncertainty

$$\sigma_R = \sqrt{\frac{1 - R_{\text{comp}}^2}{N}}.$$

Fig. 5 below shows the recovered clocks for an example of 5 datasets used to perform the alignment.

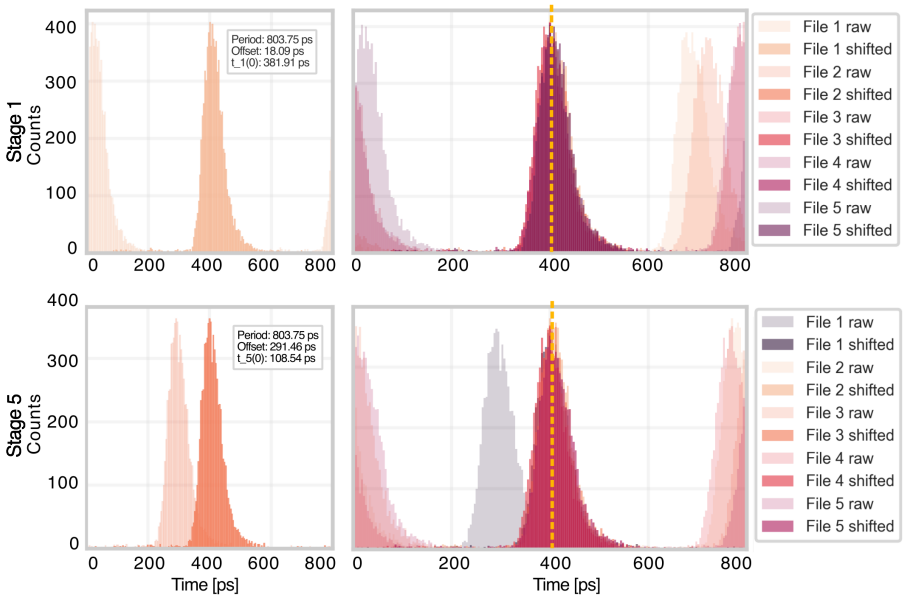

FIG. 5: **Clock recovery and temporal alignment.** (Left) Detection histograms modulo the recovered pump period for Stage 1 and Stage 5, showing well-localized arrival windows. (Right) Multiple acquisitions aligned to a common reference point within the recovered period, enabling pulse-by-pulse identification across stages.

Applying this method to the experimental data yields

$$R_{\text{comp}} = 0.93 \pm 0.01,$$

demonstrating that the two observables exhibit a high degree of practical compatibility.

SUPPLEMENTARY SECTION V: QUANTUM RANDOM NUMBERS EXTRACTION PROCEDURE

In our experiment the trusted receiver has three spatial modes monitored by single–photon detectors, and in an accepted round exactly one of these detectors clicks, so the physicaly outcome in round t can be written as

$$T_t \in \{1, 2, 3\}, \tag{8}$$

where $T_t = m$ means that the unique click occurred in spatial mode m. In addition, the classical setting (KBCS context) is

$$X_t \in \{1, \ldots, 5\}, \tag{9}$$

chosen uniformly at random before the detection event, so that the raw laboratory record is the ternary string

$$(X_t, T_t)_{t=1}^{n_{\text{det}}}. \tag{10}$$

For the security proof we do not treat T_t as the final random variable; instead, for each context i we fix a deterministic rule that converts the single click into a pair of eigenvalues $(A_i, A_{i+1}) \in \{\pm 1\}^2$ of the KCBS observables. Denote by m_i, m_{i+1} and m_{aux} the three spatial modes used in context i (two KCBS modes and one auxiliary mode). The deterministic classical map

$$f_i : \{m_i, m_{i+1}, m_{\text{aux}}\} \longrightarrow \{\pm 1\}^2 \tag{11}$$

is chosen as

$$\begin{aligned} T_t = m_i &\;\Rightarrow\; (A_i, A_{i+1}) = (+1, -1), \\ T_t = m_{i+1} &\;\Rightarrow\; (A_i, A_{i+1}) = (-1, +1), \qquad \text{when } X_t = i, \\ T_t = m_{\text{aux}} &\;\Rightarrow\; (A_i, A_{i+1}) = (-1, -1), \end{aligned} \tag{12}$$

and the analogous rule is used for the context (A_5, A_1'). Therefore every accepted ternary outcome is mapped to a unique dichotomic pair, and all three detectors appear explicitly in (12). Only events with no click or multiclick are rejected and never enter the data set. From (12) the single-body and two-body correlators used in the SDP are obtained directly from the ternary frequencies. Writing

$$p_{m|i} := \Pr(T_t = m \mid X_t = i, \text{click}_t = 1), \qquad m \in \{m_i, m_{i+1}, m_{\text{aux}}\}, \tag{13}$$

we have, for example,

$$\begin{aligned} \langle A_i \rangle &= (+1)\, p_{m_i|i} + (-1)\, p_{m_{i+1}|i} + (-1)\, p_{m_{\text{aux}}|i} = p_{m_i|i} - p_{m_{i+1}|i} - p_{m_{\text{aux}}|i}, \\ \langle A_{i+1} \rangle &= (-1)\, p_{m_i|i} + (+1)\, p_{m_{i+1}|i} + (-1)\, p_{m_{\text{aux}}|i} = -p_{m_i|i} + p_{m_{i+1}|i} - p_{m_{\text{aux}}|i}, \\ \langle A_i A_{i+1} \rangle &= (-1)\, p_{m_i|i} + (-1)\, p_{m_{i+1}|i} + (+1)\, p_{m_{\text{aux}}|i} = -p_{m_i|i} - p_{m_{i+1}|i} + p_{m_{\text{aux}}|i}. \end{aligned} \tag{14}$$

In other words, the three detectors contribute to the KCBS expectation values through different linear combinations of their conditional click probabilities, and these expectation values are exactly the quantities that enter the moment constraints of the SDP via

$$\Gamma_{A_i, \mathbb{I}} = \langle A_i \rangle, \qquad \Gamma_{A_i, A_{i+1}} = \langle A_i A_{i+1} \rangle, \tag{15}$$

and similarly for all i and for the pair (A_5, A_1'). Hence the security proof does not ignore any detector; it simply re-expresses the ternary statistics in terms of the dichotomic KCBS observables. Once the dichotomic observable A_i is fixed in this way, the raw bit for round t is defined by the map

$$A_t := \frac{1 - A_{X_t}}{2} \in \{0, 1\}. \tag{16}$$

The adversary learns the context X_t and holds a quantum system E_t correlated with the accepted state $\rho_{AE}^{(1)}$. For the security analysis the relevant random variable is therefore (A_t, X_t), not (T_t, X_t). The guessing probability that we bound in the SDP is

$$P_{\text{guess}}(A|EX) \;=\; \max_{\{E_x\}} \sum_{x \in \{1, \ldots, 5\}} p_X(x) \sum_{a \in \{0,1\}} \Pr(A = a \mid X = x)\; \text{tr}\big[\rho_{E|a,x}^{(1)} M_a^{(x)}\big], \tag{17}$$

with dichotomic Eve operators E_x as in the main text, and the single–round min–entropy is

$$H_{\min}(A \mid EX) = -\log_2 P^{\max}_{\text{guess}}(A|EX), \tag{18}$$

where $P^{\max}_{\text{guess}}(A|EX)$ is realized by the optimal moment matrix $\Gamma^\star$ solving the SDP. The whole semidefinite program is written in terms of the operators

$$\mathcal{M} = \{\mathbb{I}, A_1, \ldots, A_5, A_1', E_1, \ldots, E_5,\ A_iA_{i+1},\ A_iE_j,\ A_iA_{i+1}E_j\}, \tag{19}$$

and the entries of Γ are of the form $\Gamma_{X,Y} = \mathrm{tr}\Big[\rho^{(1)}_{AE} X^\dagger Y\Big]$, therefore the optimisation problem is posed entirely for the dichotomic observables and their products. The ternary nature of the hardware enters only through the numerical values of the correlators (14); there is no separate variable in the SDP representing T_t.

It is important to understand why the deterministic classical map from (T_t, X_t) to (A_t, X_t) does not compromise security even though it can introduce additional structure in the binary string. Let

$$B_t := (T_t, X_t), \qquad A_t = g(B_t) \tag{20}$$

with g the fixed deterministic post–processing defined by (12) and (16). For any side information E held by Eve and any deterministic function g, the optimal guessing probabilities satisfy

$$P_{\text{guess}}(A|EX) \ \geq\ P_{\text{guess}}(B|EX), \tag{21}$$

because Eve can always simulate the guess of B by first guessing A and then applying g in her classical post–processing, but not conversely. Consequently,

$$H_{\min}(A \mid EX) \ \leq\ H_{\min}(B \mid EX), \tag{22}$$

so the min–entropy of the binary variable is always smaller than or equal to the min–entropy of the full ternary record. Therefore using $H_{\min}(A \mid EX)$ for randomness extraction is conservative: any loss of entropy caused by the ternary→binary map is automatically taken into account because the SDP directly upper bounds $P_{\text{guess}}(A|EX)$ after this map. No new correlation with Eve is created by the post–processing; all correlations are already present in $\rho^{(1)}_{AE}$, and the optimisation over Eve's measurements in (17) fully exploits them.

In practice one may still store the experimental data in a ternary format (for instance, as labels $\{1, 2, 3\}$ for the three detectors), but for the security proof there is no need to certify a ternary min–entropy. The laboratory string is first converted deterministically to the KCBS outcomes through (12), from which the correlators $\langle A_i\rangle$ and $\langle A_iA_{i+1}\rangle$ and the KCBS value are computed; then the same map (16) defines the binary raw bit A_t, and the SDP plus entropy–accumulation bound give a certified rate

$$H^{\varepsilon_{\text{tot}}}_{\min}(A^{n_{\text{det}}} \mid E^{n_{\text{det}}} X^{n_{\text{det}}}) \geq n_{\text{det}}\, h(\hat{\chi}_{\text{KCBS}} + \Delta, \mathbf{w}, \delta, R) - \beta\sqrt{n_{\text{det}}}, \tag{23}$$

which is then fed into the Toeplitz extractor. All three detectors are thus used twice: first in the construction of the KCBS witness and the moment matrix through (14), and second in the definition of the actual binary variable A_t whose entropy is being certified and eventually hashed.

SUPPLEMENTARY SECTION VI: NIST RANDOMNESS TESTS

After the randomness extraction procedure, the randomness of the extracted sequence was confirmed by performing statistical tests by the NIST SP 800-22 suite [14] . Due to the modest length of the sequences (N = 65,100 bits), only a restricted subset of the tests was applied. In particular, we considered tests that probe global bias, short-range correlations, run structure, and local entropy, which remain meaningful at this sequence length. Tests requiring asymptotic statistics or multiple long sequences were excluded. The results, summarized in Table II, indicate that the extracted bit string is compatible with an ideal sequence of independent and identically distributed bits, within the bounds set by the available data length.

Test	Parameter	Result
Monobit frequency	–	$p = 0.975$
Block frequency	$M = 100$	$p = 0.287$
Runs test	–	$p = 0.351$
Serial test	$m = 2$	$p = 0.117$
Approximate entropy	$m = 2$	0.6931
Longest run of ones	–	15

TABLE II: Results of statistical consistency tests applied to a Toeplitz-extracted bit sequence of length 65,100. Only tests that remain meaningful at this sequence length were considered. The reported values are consistent with those expected for a uniform random sequence, within finite-size effects.

SUPPLEMENTARY SECTION VII: SEMI-DI QRNG SECURITY PROOF

Security Proof

In the semi-device-independent protocol considered, the quantum source is treated as untrusted, while the measurement module, the single-click acceptance rule, and the setting-choice randomness are trusted, and a protocol round is accepted only if the trusted detectors collectively register a single click across the three signal modes, while any no-click or multi-click event results in abortion of the round. The accepted-round event is modeled by an acceptance effect (not a projector): operationally it corresponds to a single registered click, while residual multiphoton contamination is not assumed away but is bounded separately through $\varepsilon_{\mathrm{mp}}$. Because physical loss precedes detection and acceptance is defined by a single registered click, we define the valid-round acceptance effect via the loss adjoint applied to the click POVM; single-photon leakage is handled operationally via acceptance and a measured bound:

$$\begin{aligned}
\Pi_{n=1}^{(\mathrm{signal})} &:= \sum_{k=1}^{3} \big(|1\rangle_k\langle 1|_k \otimes \prod_{j\neq k} |0\rangle_j\langle 0|_j \big), \\
E_{\mathrm{val}} &:= \mathcal{L}_{\boldsymbol{\eta}}^{\dagger}\Big(\sum_{s=1}^{3} C_s \Big), \\
&\qquad \text{with} \quad 0 \preceq C_s \preceq \mathbb{I}, \quad C_{\mathrm{nc}}, C_{\geq 2} \succeq 0, \\
&\qquad \sum_{s=1}^{3} C_s + C_{\mathrm{nc}} + C_{\geq 2} = \mathbb{I}, \quad 0 \preceq E_{\mathrm{val}} \preceq \mathbb{I}, \\
\mathcal{E}_{\mathrm{post}}[\sigma] &:= \sqrt{E_{\mathrm{val}}}\, \sigma \, \sqrt{E_{\mathrm{val}}}, \\
\rho_{AE}^{(1)} &:= \frac{\mathcal{E}_{\mathrm{post}}\big[\tilde{\rho}_{AE}\big]}{\mathrm{tr}\big[E_{\mathrm{val}}\tilde{\rho}_{AE}\big]}.
\end{aligned} \tag{24}$$

POVM conditions are

$$C_{\mathrm{nc}}: \text{ no-click}, \qquad C_{\geq 2}: \text{ multi-click}, \qquad \sum_{s=1}^{3} C_s: \text{ single-click acceptance}. \tag{25}$$

The operator $\Pi_{n=1}^{(\mathrm{signal})}$ is used only to quantify single-photon leakage, while E_{val} enforces acceptance (a single registered click). The conditional state is defined by acceptance via E_{val}, and the residual multi-photon contribution is handled through the bound $\varepsilon_{\mathrm{mp}}$, thereby removing any assumption on the dimensionality of Eve's Hilbert space. Multiphoton leakage bound is enforced as

$$\mathrm{tr}\big[(\mathbb{I} - \Pi_{n=1}^{(\mathrm{signal})})\, \mathcal{L}_{\boldsymbol{\eta}}(\tilde{\rho}_A)\big] \leq \varepsilon_{\mathrm{mp}}. \tag{26}$$

(e.g., from a measured heralded $g^{(2)}(0)$), here $\tilde{\rho}_A = \mathrm{tr}_E[\tilde{\rho}_{AE}]$. We fold this bounded leakage into a conservative robustness penalty for the guessing probability:

$$\varepsilon_{\mathrm{acc}} := \min\left\{1,\ \frac{\varepsilon_{\mathrm{mp}}}{p_{\mathrm{acc}}}\right\}, \qquad P_{\mathrm{guess}}^{\mathrm{max,rob}}(k, \mathbf{w}, \delta, R) \ := \ (1 - \varepsilon_{\mathrm{acc}})\, P_{\mathrm{guess}}^{\mathrm{max}}(k, \mathbf{w}, \delta, R) \ + \ \varepsilon_{\mathrm{acc}}. \tag{27}$$

Eve prepares an arbitrary state $\tilde{\rho}_{AE}$ on $\mathcal{H}_A^{\mathrm{phys}} \otimes \mathcal{H}_E$. The post-selection map is the Lüders instrument defined above. This post-selection map is completely positive and trace–non–increasing; the normalization produces a valid density operator. Here C_s are click effects (POVM elements) at the detection stage (post-loss), satisfying $0 \preceq C_s \preceq \mathbb{I}$ and $\sum_{s=1}^{3} C_s \preceq \mathbb{I}$. The map $\mathcal{L}_{\boldsymbol{\eta}}$ is a trace-preserving loss channel acting on A and is used via its adjoint in E_{val}.

$$\mathcal{L}_{\boldsymbol{\eta}} = \bigotimes_{s=1}^{3} \mathcal{L}_{\eta_s}, \qquad \mathcal{L}_{\eta_s}[\rho] = \eta_s\, \rho + (1 - \eta_s)\, \mathrm{tr}[\rho]\, |0\rangle\langle 0|, \qquad \mathcal{L}_{\eta_s}^{\dagger}[M] = \eta_s\, M + (1 - \eta_s)\langle 0|M|0\rangle\, \mathbb{I}. \tag{28}$$

The measurement section is trusted and implements six fixed dichotomic observables $A_1, \ldots, A_5, A_1'$ with eigenvalues ± 1. Approximate compatibility of adjacent pairs is enforced via calibrated near-commutation; their imperfect overlap

is characterized solely by the measured correlation $R := \langle A_1 A_1' \rangle_{\text{obs}}$ with $|R| \leq 1$. The three spatial modes have detection efficiencies $\eta_1, \eta_2, \eta_3 \in (0, 1]$. The mapping $\pi : \{1, \ldots, 5\} \to \{1, 2, 3\}$ between contexts and spatial modes is defined as: $\pi(1) = \pi(4) = 1$ (contexts 1 and 4 use mode 1); $\pi(2) = \pi(5) = 2$ (contexts 2 and 5 use mode 2); $\pi(3) = 3$ (context 3 uses mode 3), giving context-dependent acceptance probabilities $p(\text{click} \mid X = i)$. Abort unless $p_{\text{acc}} \geq \eta_{\min}$ (run criterion only; not an SDP constraint).

Define

$$\begin{aligned} p_{\text{acc}} &:= \frac{1}{5} \sum_{i=1}^{5} p(\text{click} \mid X = i), \qquad w_i := \frac{p(\text{click} \mid X = i)}{5\, p_{\text{acc}}}, \\ &\sum_{i=1}^{5} w_i = 1, \qquad \underline{w}_i \leq w_i \leq \overline{w}_i \quad \text{(Hoeffding per context [15])}. \end{aligned} \tag{29}$$

Note that by construction we have $\sum_{i=1}^{5} w_i = 1$. All per-accepted-round statements depend on $\mathbf{w} = (w_1, \ldots, w_5)$. The observed marginal biases satisfy $|\langle A_i \rangle_{\text{obs}}| \leq \delta$, where δ is a predetermined bias limit set by device calibration (e.g. the maximum $|\langle A_i \rangle|$ observed during calibration, expanded by a small statistical margin).

The label $X \in \{1, \ldots, 5\}$ is drawn uniformly from a trusted classical randomness source that remains private to Alice until the corresponding outcome is recorded, and after data collection, the public Toeplitz seed S used for extraction is generated independently of (A^n, E^n, X^n). The proof therefore certifies private randomness within the stated semi-device-independent trust model; it is not a net randomness-expansion claim, since the setting variable X is consumed from an independent trusted seed and the certified quantity is conditioned on X and on acceptance.

The accepted rounds are treated within the entropy-accumulation framework; concentration of empirical statistics uses Hoeffding bounds per context and a union bound. The measurement process in round t is described by the quantum channel

$$\mathcal{N}_{X_t}[\rho_{AE}] = \sum_{a_i, a_{i+1} \in \{\pm 1\}} \sum_{c \in \{0,1\}} \text{tr}_A \left[(N^{(i)}_{a_i, a_{i+1}, c} \otimes \mathbb{I}_E) \rho_{AE} \right] |a_i, a_{i+1}, c\rangle\langle a_i, a_{i+1}, c|, \tag{30}$$

Here, $c \in \{0, 1\}$ flags rejection or acceptance, and $N^{(i)}$ specifies the corresponding instrument.

$$N^{(i)}_{a_i, a_{i+1}, 1} := \sqrt{E_{\text{val}}}\, \Pi^{(i)}_{a_i, a_{i+1}} \sqrt{E_{\text{val}}}, \qquad N^{(i)}_{\cdot, \cdot, 0} := \mathbb{I}_A - \sum_{a_i, a_{i+1}} N^{(i)}_{a_i, a_{i+1}, 1}, \qquad \sum_{a_i, a_{i+1}, c} N^{(i)}_{a_i, a_{i+1}, c} = \mathbb{I}_A. \tag{31}$$

This channel formalism is essential for applying the entropy accumulation theorem.

$$\Pi^{(i)}_{a_i, a_{i+1}} := \frac{\mathbb{I} + a_i A_i}{2} \frac{\mathbb{I} + a_{i+1} A_{i+1}}{2}, \qquad a_i, a_{i+1} \in \{\pm 1\}, \quad \text{with } A_{i+1} \equiv A_1' \text{ when } i = 5. \tag{32}$$

For $\epsilon_{\text{com}} > 0$ the product form above is a shorthand for the ideal commuting case; operationally, the channel is implemented by a jointly measurable POVM with effects $M^{(i)}_{a_i, a_{i+1}} \succeq 0$ on A such that the observed moments coincide with those of the formal expressions. The SDP uses only the moment constraints (including near-commutation), not the explicit product form, so complete positivity is guaranteed by Naimark dilation without imposing further algebraic identities.

The post-selected joint system satisfies

$$\rho^{(1)}_{AE} \in \mathcal{B}(\mathbb{C}^3 \otimes \mathcal{H}_E), \qquad \rho^{(1)}_{AE} \succeq 0, \qquad \text{tr}\, \rho^{(1)}_{AE} = 1. \tag{33}$$

The post-selection map $\sigma \mapsto E_{\text{val}}^{1/2} \sigma E_{\text{val}}^{1/2}$ is trace–non–increasing; normalization by $p = \text{tr}[E_{\text{val}} \tilde{\rho}_{AE}]$ yields the conditional state $\rho^{(1)}_{AE}$. With observables

$$A_1,\ A_2,\ A_3,\ A_4,\ A_5,\ A_1', \qquad A_k^2 = \mathbb{I}, \quad (A_1')^2 = \mathbb{I}, \tag{34}$$

We also impose Hermiticity and ± 1 spectra for Alice's and Eve's observables:

$$A_i^\dagger = A_i, \quad (A_1')^\dagger = A_1', \quad E_i^\dagger = E_i, \quad E_i^2 = \mathbb{I}_E. \tag{35}$$

Approximate compatibility of nominally commuting pairs is enforced via the calibrated near-commutation constraints in (50)–(52); no exact commutators are assumed.

The imperfect compatibility between A_1 and A_1' is quantified experimentally by measuring their correlation:

$$R := \langle A_1 A_1' \rangle_{\text{obs}}. \tag{36}$$

This measured value, along with its confidence interval, is incorporated directly into the semidefinite program through the constraint

$$\underline{R} \leq \Gamma_{A_1, A_1'} \leq \overline{R}, \tag{37}$$

where $[\underline{R}, \overline{R}]$ is determined from experimental data with statistical uncertainty. No specific algebraic form for A_1' is assumed; the SDP approach handles any physical implementation that produces the observed correlation.

Projective measurements on a multi-rail qutrit encode the logical basis states $|0\rangle, |1\rangle, |2\rangle$ in three spatial waveguides. The KCBS observables are defined as

$$A_k = 2|v_k\rangle\langle v_k| - \mathbb{I}_{\mathcal{H}_A}, \quad k = 1, \ldots, 5, \tag{38}$$

where the vectors $|v_k\rangle$ satisfy the pentagram relations

$$\langle v_k | v_{k+1} \rangle = 0 \quad \text{(indices modulo 5)}, \quad \sum_{k=1}^{5} |v_k\rangle\langle v_k| \neq \mathbb{I}_{\mathcal{H}_A}. \tag{39}$$

These relations define the ideal target observables ($A_k^2 = \mathbb{I}$); in the proof we enforce approximate compatibility for adjacent pairs via (50).

The bias bound is defined by

$$\delta_k := \sqrt{\frac{\ln(2/\varepsilon_{\text{cal}})}{2n_k}}, \quad \langle A_k \rangle \in [\langle A_k \rangle_{\text{obs}} - \delta_k, \langle A_k \rangle_{\text{obs}} + \delta_k] \tag{40}$$

For implementation we use the uniform single-body bound $\delta := \max_k \delta_k$, i.e., $|\Gamma_{A_k, \mathbb{I}}| \leq \delta$, $|\Gamma_{A_1', \mathbb{I}}| \leq \delta$. For convenience we write the single-body expectation intervals as

$$\underline{b}_k := \langle A_k \rangle_{\text{obs}} - \delta_k, \qquad \overline{b}_k := \langle A_k \rangle_{\text{obs}} + \delta_k. \tag{41}$$

Now we can define the KCBS combination

$$\chi_{\text{KCBS}} := \langle A_1 A_2 \rangle + \langle A_2 A_3 \rangle + \langle A_3 A_4 \rangle + \langle A_4 A_5 \rangle + \langle A_5 A_1' \rangle. \tag{42}$$

with

$$\chi_{\text{KCBS}}^{\text{NCHV}} \geq -3, \qquad \chi_{\text{KCBS}}^{\text{QM}} \geq 5 - 4\sqrt{5} \approx -3.944. \tag{43}$$

Given a uniformly random $X \in \{1, \ldots, 5\}$, if $X = i$, Alice measures (A_i, A_{i+1}) and outputs a raw bit defined by

$$A_t := \frac{1 - A_{X_t}}{2} \in \{0, 1\}. \tag{44}$$

The certified raw variable is therefore the binary string $A^{n_{\text{det}}}$ obtained from the operator-defined outcomes A_{X_t}; any laboratory ternary detector label is only an auxiliary record and is not the variable to which the min-entropy bound is applied. while after learning $X = i$, Eve performs a dichotomic measurement E_i on her subsystem characterized by

$$-\mathbb{I}_E \preceq E_i \preceq \mathbb{I}_E, \quad E_i^\dagger = E_i, \qquad \langle A_i \otimes E_i \rangle = \text{tr}\Big(\rho_{AE}^{(1)}(A_i \otimes E_i)\Big). \tag{45}$$

Note that the expectation is taken with respect to the post-selected state $\rho_{AE}^{(1)}$. Eve's observable E_i is chosen after learning $X = i$ (within-round adaptivity); across rounds, the EAT Markov premises ensure no signaling. Since Alice's raw outcome is binary, the Helstrom theorem guarantees that a two-outcome POVM is optimal for minimum-error discrimination, so restricting Eve to dichotomic measurements $\{E_i, \mathbb{I}_E - E_i\}$ incurs no loss of generality. For an accepted round where context i is measured, the conditional success probability for Eve to guess Alice's bit is

$$P_{\text{succ}}^{(i)} = \frac{1}{2}\left(1 + \langle A_i \otimes E_i \rangle\right). \tag{46}$$

and averaging over the five contexts with weights w_i yields

$$P_{\text{succ}}^{\text{det}} = \sum_{i=1}^{5} w_i \, \frac{1 + \langle A_i \otimes E_i \rangle_{\rho_{AE}^{(1)}}}{2}. \tag{47}$$

Since no-click and multi-click events are discarded, all security statements condition on accepted detected rounds, making the per-accepted-round guessing probability exactly the success probability

$$P_{\text{guess}} = P_{\text{succ}}^{\text{det}} = \sum_{i=1}^{5} w_i \, \frac{1 + \langle A_i \otimes E_i \rangle_{\rho_{AE}^{(1)}}}{2}. \tag{48}$$

where the expectation is taken with respect to the post-selected state $\rho_{AE}^{(1)}$.

We define the operator set

$$\begin{aligned} \mathcal{M} =& \{\mathbb{I}, A_1, \ldots, A_5, A_1', E_1, \ldots, E_5, A_1A_2, A_2A_3, A_3A_4, A_4A_5, A_5A_1'\} \cup \\ & \{A_iE_j : \ 1 \le i, j \le 5\} \ \cup \ \{A_iA_{i+1}E_j : \ 1 \le i \le 4, \ 1 \le j \le 5\} \ \cup \ \{A_5A_1'E_j : \ 1 \le j \le 5\}. \end{aligned} \tag{49}$$

On the operator family (49) we impose normalization and algebraic constraints that encode the intended measurement structure. In particular, approximate compatibility of the adjacent KCBS settings is enforced at the level of the moment matrix by requiring near-commutation of nominally commuting pairs as follows:

$$\begin{aligned} & \big\| \, [X, Y] \, \big\|_{\Gamma} := \sqrt{\Gamma_{XY,XY} - 2\,\Gamma_{XY,YX} + \Gamma_{YX,YX}} \ \le \ \epsilon_{\text{com}}, \\ & (X, Y) \in \{(A_1, A_2), (A_2, A_3), (A_3, A_4), (A_4, A_5), (A_5, A_1')\}. \end{aligned} \tag{50}$$

We work at level-2+AB over (49), including all A_iE_j cross terms and the full KCBS block $A_1A_2, \ldots, A_4A_5, A_5A_1'$. Near-commutation is enforced via (50); we do not impose exact anticommutator (Jordan) identities when $\epsilon_{\text{com}} > 0$.

A dedicated detector-tomography run yields

$$\epsilon_{\text{com}}^{\text{cal}} = 0.047 \pm 0.003, \tag{51}$$

which is substituted for ϵ_{com} in every semidefinite program call, while the commutator norm is defined as

$$\epsilon_{\text{com}} := \epsilon_{\Gamma}^{\text{cal}}, \qquad \text{where } \epsilon_{\Gamma}^{\text{cal}} \text{ is obtained from the calibrated Gram-moment norm } |[A_i, A_{i+1}]|_{\Gamma}. \tag{52}$$

For every compatible pair (i, j) we compute the empirical correlator $C_{ij}^{\text{obs}} = \frac{1}{n_{ij}} \sum_{t \in \mathcal{D}_{ij}} a_t^{(i)} a_t^{(j)}$, on the n_{ij} detected rounds $\mathcal{D}_{ij}$ in which both observables are measured, and using Hoeffding's inequality with confidence level $1 - \varepsilon_{\text{cal}}$, pass the interval $\left[C_{ij}^{\text{obs}} - \sqrt{\frac{\ln(2/\varepsilon_{\text{cal}})}{2\,n_{ij}}}, \quad C_{ij}^{\text{obs}} + \sqrt{\frac{\ln(2/\varepsilon_{\text{cal}})}{2\,n_{ij}}}\right]$ directly into the bounds $\underline{C}_{ij}, \overline{C}_{ij}$ of Eq. (20). Experimental data impose interval constraints:

$$\begin{aligned} & \Gamma_{\mathbb{I},\mathbb{I}} = 1, \qquad \underline{b}_k \le \Gamma_{A_k,\mathbb{I}} \le \bar{b}_k, \\ & \underline{C}_{ij} \le \Gamma_{A_i,A_j} \le \overline{C}_{ij} \ (i \sim j), \qquad R^{\min} \le \Gamma_{A_1,A_1'} \le R^{\max}, \\ & \Gamma_{E_i,E_i} = 1 \quad (i = 1, \ldots, 5), \qquad [A_i \otimes \mathbb{I}_E, \ \mathbb{I}_A \otimes E_j] = 0 \ \ \forall i, j. \end{aligned} \tag{53}$$

Here $i \sim j$ denotes adjacent (KCBS-compatible) pairs.

By Hermiticity, $\Gamma_{E_i,E_j} = \overline{\Gamma_{E_j,E_i}}$ for all i, j. We also set $R^{\min} = R - \sqrt{\frac{\ln(2/\varepsilon_{\text{cal}})}{2n_R}}$ and $R^{\max} = R + \sqrt{\frac{\ln(2/\varepsilon_{\text{cal}})}{2n_R}}$, where n_R is the number of accepted rounds used to estimate R. In every round, the acceptance effect E_{val} is applied before data are recorded, so Eve's state and the SDP refer to the conditional state $\rho_{AE}^{(1)}$ defined in Eq. (33). No terms involving $\Gamma_{X,\mathbb{I}}$ with $X = E_{\text{val}}$ appear, because acceptance is enforced by normalization.

Since Eve's goal is to maximize her guessing probability conditioned on detection events, the objective function must account for the non-uniform detection weights w_i. The optimization problem becomes:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{5} \bar{w}_i \, \frac{1 + \Gamma_{A_i,E_i}}{2} \\ \text{subject to} \quad & \underline{w}_i \le \bar{w}_i \le \overline{w}_i, \ \sum_{i=1}^{5} \bar{w}_i = 1, \\ & \Gamma \succeq 0, \ \text{constraints (34), (35), (50), (53).} \end{aligned} \tag{54}$$

Here, the variables $\bar{w}_i$ are the SDP decision weights constrained by the empirical intervals $[\underline{w}_i, \overline{w}_i]$ and the normalization $\sum_i \bar{w}_i = 1$, while the witness $\hat{\chi}_{\mathrm{KCBS}}$ itself uses uniform context sampling.

Note that the contextuality witness and its estimator use uniform context weights; only the guessing objective employs the random post-selection weights $\mathbf{w}$ (via worst-case intervals). Since the objective function and feasible set are convex, Eq. (54) constitutes a semidefinite program whose optimum is denoted by $P^{\max}_{\mathrm{succ}}$.

Theorem. Any extreme optimal $\Gamma^\star$ satisfies $r(r+1)/2 \le m_{\mathrm{act}}$ (Pataki), where $r = \operatorname{rank}\Gamma^\star$ and m_{act} is the number of active affine constraints at optimum. For our data sets, numerical solutions give $r \le 4$, which shows that an ancilla of dimension 4 is *sufficient* for those instances.

Proof. Pataki's theorem applies directly. The removal of detection-operator constraints reduces m relative to earlier counts; for the reported data instances, the observed optimum satisfies $r \le 4$. By GNS/Stinespring, the corresponding moments are realized on $\mathbb{C}^3 \otimes \mathbb{C}^4$, which attains $F(\Gamma^\star)$ for those instances. □

Optimizing the objective in Eq. (54) yields

$$P^{\max}_{\mathrm{succ}} = \max \sum_{i=1}^{5} \bar{w}_i \, \frac{1 + \Gamma_{A_i,E_i}}{2}, \qquad P^{\max,\mathrm{rob}}_{\mathrm{guess}} = (1 - \varepsilon_{\mathrm{acc}})\, P^{\max}_{\mathrm{succ}} + \varepsilon_{\mathrm{acc}} \, . \tag{55}$$

The single-round conditional min-entropy is then bounded by

$$H_{\min}(A|EX, \mathrm{click} = 1) \;\ge\; -\log_2 P^{\max,\mathrm{rob}}_{\mathrm{guess}}(k, \mathbf{w}, \delta, R) \;=:\; h(k, \mathbf{w}, \delta, R). \tag{56}$$

The quantity $h(k, \mathbf{w}, \delta, R)$ is the certified single-round entropy rate used in the finite-size analysis. Relations (53)–(56) form a complete, fully algebraic pipeline aligned with the numerical implementation: inserting the experimentally determined intervals $\{\underline{b}_k, \overline{b}_k, \underline{C}_{ij}, \overline{C}_{ij}, R^{\min}, R^{\max}\}$ into constraints (53) and solving the SDP (54) outputs $P^{\max}_{\mathrm{succ}}$; Eq. (55) converts this into $P^{\max,\mathrm{rob}}_{\mathrm{guess}}$; finally, Eq. (56) yields the provable amount of certified private randomness extractable from a single accepted round.

Considering finite-size corrections, we analyze the accepted rounds (in total n_{det}), where in each accepted round the corresponding compatible pair is measured, and the single bit $A_t = \frac{1-A_{X_t}}{2} \in \{0, 1\}$ is recorded. The single-round contextual estimator is defined by

$$Z_t = A_1 A_2 \, \mathbf{1}_{\{X_t=1\}} + A_2 A_3 \, \mathbf{1}_{\{X_t=2\}} + A_3 A_4 \, \mathbf{1}_{\{X_t=3\}} + A_4 A_5 \, \mathbf{1}_{\{X_t=4\}} + A_5 A'_1 \, \mathbf{1}_{\{X_t=5\}}, \qquad |Z_t| \le 1, \tag{57}$$

and its experimental average is defined as follows

$$\hat{C}_{i,i+1} := \frac{1}{n_i} \sum_{\substack{t:\, X_t = i \\ \mathrm{click}_t = 1}} a_t^{(i)} \, a_t^{(i+1)}, \qquad n_i := \big|\{t :\, X_t = i,\ \mathrm{click}_t = 1\}\big|, \qquad \hat{\chi}_{\mathrm{KCBS}} := \sum_{i=1}^{5} \hat{C}_{i,i+1}. \tag{58}$$

Denote by $\chi_{\mathrm{KCBS}_{\mathrm{true}}} = \mathbb{E}[Z_t]$ the true KCBS value of the underlying distribution. Hoeffding's inequality applied per context and a union bound over the five pairs yield the weighted deviation term

$$\Delta := \sum_{i=1}^{5} \sqrt{\frac{\ln\!\big(10/\varepsilon_{\mathrm{est}}\big)}{2\, n_i}}. \tag{59}$$

Except with probability $\varepsilon_{\mathrm{est}}$, we then have

$$\chi_{\mathrm{KCBS}_{\mathrm{true}}} \le \hat{\chi}_{\mathrm{KCBS}} + \Delta. \tag{60}$$

All entropies are conditioned on acceptance:

$$H_{\min}(A_t \mid E_t, X_t, \mathrm{click}_t{=}1) \ge -\log_2 P^{\max,\mathrm{rob}}_{\mathrm{guess}}\big(\hat{\chi}_{\mathrm{KCBS}} + \Delta,\ \mathbf{w},\ \delta,\ R\big) = h\big(\hat{\chi}_{\mathrm{KCBS}} + \Delta, \mathbf{w}, \delta, R\big) \tag{61}$$

For any value k consistent with the constraints (53) and the near-commutation bounds, the semidefinite program in Eq. (62) returns the maximal guessing probability

$$\begin{aligned} P^{\max}_{\mathrm{guess}}(k; \mathbf{w}, \delta, R) = \sup_{\rho_{AE},\, E_i} & \sum_{i=1}^{5} \bar{w}_i \cdot \frac{1 + \langle A_i \otimes E_i \rangle}{2} \\ & \text{subject to} \\ & \text{the moment matrix constraints hold and} \\ & \langle A_1 A_2 \rangle + \cdots + \langle A_4 A_5 \rangle + \langle A_5 A'_1 \rangle = k. \end{aligned} \tag{62}$$

Denote

$$h(k, \mathbf{w}, \delta, R) = -\log_2 P_{\text{guess}}^{\max,\text{rob}}(k, \mathbf{w}, \delta, R) \ . \tag{63}$$

where $P_{\text{guess}}^{\max}(k, \mathbf{w}, \delta, R)$ is the maximum guessing probability obtained from solving the SDP in Eq. (54) with KCBS value k, weights $\mathbf{w}$, bias bound δ, and overlap R. We compute single-round entropy via the SDP-defined function $h(k, \mathbf{w}, \delta, R)$ and use it within an entropy-accumulation bound below. We denote the number of accepted rounds by

$$n_{\text{det}} := \sum_{t=1}^{n} \mathbf{1}_{\{\text{click}_t=1\}}. \tag{64}$$

Each round is generated by a CPTP map $\mathcal{M}_t$ with $X_{<t} \to E_{<t} \to (A_t, E_t, X_t)$, fulfilling the EAT hypotheses.

Define the overall security parameter as $\varepsilon_{\text{tot}} = \varepsilon_{\text{est}} + \varepsilon_{\text{sm}} + \varepsilon_{\text{sdp}}$, where ε_{sm} is the EAT smoothing parameter and ε_{sdp} bounds numerical inaccuracies (primal–dual gap of the SDP solver). With probability at least $1 - \varepsilon_{\text{tot}}$, the certified smooth min-entropy reads

$$H_{\min}^{\varepsilon_{\text{tot}}}(\cdots) \ \geq \ n_{\text{det}} \cdot h\big(\hat{\chi}_{\text{KCBS}} + \Delta, \mathbf{w}, \delta, R\big) \ - \ \beta\sqrt{n_{\text{det}}}, \qquad \beta := L\sqrt{2\ln\big(1/\varepsilon_{\text{sm}}\big)}. \tag{65}$$

We bound the k-sensitivity of the single-round entropy via the dual slope:

$$L := \sup_{k\in\mathcal{I}} \left|\frac{\partial}{\partial k} h(k, \mathbf{w}, \delta, R)\right| \ = \ \frac{\nu^\star}{\ln 2} \cdot \frac{1}{\min_{k\in\mathcal{I}} P_{\text{guess}}^{\max,\text{rob}}(k)}. \tag{66}$$

where $\mathcal{I}$ denotes the confidence interval for k induced by the estimate of the KCBS value: $\mathcal{I} = \big[\hat{\chi}_{\text{KCBS}} - \Delta,\ \hat{\chi}_{\text{KCBS}} + \Delta\big] \cap [-5, 5]$, and the function h is defined by Eq. (63) and fully determined through the semidefinite program Eq. (62) once the experimentally estimated imperfections $\mathbf{w}, \delta, R$ and the empirical KCBS value $\hat{\chi}_{\text{KCBS}}$ are incorporated.

For each accepted round t, the experiment records a single context $X_t \in \{1, \ldots, 5\}$ along with the outcome pair $(a_t^{(i)}, a_t^{(i+1)})$ of the commuting observables (A_i, A_{i+1}) defining that context. Define $Z_t^{(i)} = a_t^{(i)} a_t^{(i+1)} \in \{\pm 1\}$ if $X_t = i$, and $Z_t^{(i)} = 0$ otherwise. Let $\chi_{\text{KCBS}_{\text{true}}}$ denote the true KCBS value. Using Hoeffding per context and a union bound over five pairs (two tails each) gives the following confidence intervals. Each random variable $Z_t^{(i)}$ takes values in $[-1, 1]$; for every fixed pair $(i, i+1)$ and $\varepsilon \in (0, 1)$,

$$\Pr\left[\left|\hat{C}_{i,i+1} - C_{i,i+1}^{\text{true}}\right| \leq \sqrt{\frac{\ln(2/\varepsilon)}{2\, n_i}}\right] \geq 1 - \varepsilon \quad (i = 1, \ldots, 5). \tag{67}$$

while a union bound over the five pairs yields, with error probability ε_{est},

$$\Pr[|\hat{\chi}_{\text{KCBS}} - \chi_{\text{KCBS}_{\text{true}}}| \leq \Delta] \geq 1 - \varepsilon_{\text{est}}, \qquad \Delta := \sum_{i=1}^{5} \sqrt{\frac{\ln\big(10/\varepsilon_{\text{est}}\big)}{2\, n_i}}. \tag{68}$$

For a fixed set of device parameters $(\mathbf{w}, \delta, R)$, the single-shot guessing probability obtained from the SDP in Eq. (54) is the optimal value of a linear functional over a convex feasible set that depends affinely on the true correlator $\chi_{\text{KCBS}}^{\text{true}}$, with numerical evaluation on a k-grid showing that $P_{\text{succ}}^{\max}(k, \mathbf{w}, \delta, R)$ is monotone non-decreasing; this empirical monotonicity is used to interpolate the entropy curve. A formal proof follows from standard Lagrange-duality arguments, and monotonicity can be verified experimentally by solving the SDP on a fine grid of k values. Define

$$g(k) = -\log_2 P_{\text{succ}}^{\max,\text{rob}}(k, \mathbf{w}, \delta, R), \qquad g'(k) \leq 0 \ \ (\text{dual slope } \nu^\star \geq 0). \tag{69}$$

Monotonicity. Writing the dual of (62) with Lagrange multiplier ν for the KCBS constraint $L(\Gamma) = k$ yields a dual objective of the form $\min\{\alpha + \nu\, k\}$ with $\nu^\star \geq 0$ at optimum; hence $P_{\text{succ}}^{\max}(k)$ is non-decreasing in k and $g(k)$ is non-increasing.

$$H_{\min}(A_t \mid E_t, X_t, \text{click}_t = 1) \geq h\big(\hat{\chi}_{\text{KCBS}} + \Delta, \mathbf{w}, \delta, R\big). \tag{70}$$

We generate the length-d Toeplitz seed S after the experiment is closed and all raw strings are fixed, thereby guaranteeing statistical independence from the data. The extractor is applied directly to the accepted-round binary string $\boldsymbol{A}$; no intermediate ternary-to-binary recoding is assumed anywhere in the certified pipeline. We choose a Toeplitz matrix $T \in \mathbb{F}_2^{m\times n_{\rm det}}$ whose seed $S \in \{0,1\}^{n_{\rm det}+m-1}$ is generated *after* all classical data have been fixed, and define $R_{\rm ext} = T\boldsymbol{A} \bmod 2$. The quantum-proof leftover-hash lemma states that for every $\varepsilon_{\rm ext} \in (0,1)$,

$$m = \left\lfloor n_{\rm det}\cdot h(\hat{\chi}_{\rm KCBS}+\Delta, \mathbf{w}, \delta, R) - \beta\sqrt{n_{\rm det}} - 2\log_2(1/\varepsilon_{\rm ext})\right\rfloor \tag{71}$$

where the $-\beta\sqrt{n_{\rm det}}$ term is the EAT finite-size (smoothing) correction, and the last term is the extractor security penalty.

Under this condition, the joint state satisfies

$$\left\|\rho_{R_{\rm ext}S\boldsymbol{EX}\,{\rm click}^{n_{\rm det}}} - \frac{I_{R_{\rm ext}}}{2^m}\otimes\rho_{S\boldsymbol{EX}\,{\rm click}^{n_{\rm det}}}\right\|_1 \le \varepsilon_{\rm fin}, \qquad \varepsilon_{\rm fin} = \varepsilon_{\rm est}+\varepsilon_{\rm sm}+\varepsilon_{\rm ext}+\varepsilon_{\rm sdp}+\varepsilon_{\rm mp} \tag{72}$$

where Eqs. (71)–(72) establish composable $\varepsilon_{\rm fin}$-randomness of the extracted bit string $R_{\rm ext}$ against arbitrary quantum adversaries. This establishes composable $\varepsilon_{\rm fin}$-security, formally stated as

$$\|\rho_{R_{\rm ext}E^nX^n} - \tau_{R_{\rm ext}}\otimes\rho_{E^nX^n}\|_{\rm tr} \le \varepsilon_{\rm fin}, \quad \tau_{R_{\rm ext}} = \mathbb{I}_{R_{\rm ext}}/2^m, \tag{73}$$

where τ_R is the uniform distribution on m bits. This trace distance bound ensures the extracted randomness is indistinguishable from uniform against any quantum adversary. Explicitly,

$$\begin{aligned}
&\text{input: } n_{\rm det},\ \hat{\chi}_{\rm KCBS},\ \mathbf{w},\ \delta,\ R,\ \varepsilon_{\rm est},\ \varepsilon_{\rm sm},\ \varepsilon_{\rm ext},\ \varepsilon_{\rm sdp},\ \varepsilon_{\rm mp}\\
&\text{certified raw string: } \boldsymbol{A}\in\{0,1\}^{n_{\rm det}} \text{ with } A_t = (1-A_{X_t})/2,\ \text{conditioned on click}_t = 1\\
&\Delta = \sum_{i=1}^{5}\sqrt{\frac{\ln(10/\varepsilon_{\rm est})}{2n_i}}, \qquad \lambda = h\big(\hat{\chi}_{\rm KCBS}+\Delta,\ \mathbf{w},\ \delta,\ R\big),\\
&\Lambda = n_{\rm det}\lambda - \beta\sqrt{n_{\rm det}}, \quad \beta = L\sqrt{2\ln(1/\varepsilon_{\rm sm})},\\
&m = \lfloor \Lambda - 2\log_2(1/\varepsilon_{\rm ext})\rfloor, \quad d = n_{\rm det}+m-1,\\
&R_{\rm ext} = T\boldsymbol{A} \bmod 2, \qquad T\in\mathbb{F}_2^{m\times n_{\rm det}} \text{ Toeplitz, seed } S\in\{0,1\}^d,\\
&\left\|\rho_{R_{\rm ext}S\boldsymbol{EX}\,{\rm click}^{n_{\rm det}}} - \frac{I_{R_{\rm ext}}}{2^m}\otimes\rho_{S\boldsymbol{EX}\,{\rm click}^{n_{\rm det}}}\right\|_1 \le \varepsilon_{\rm fin},\\
&\varepsilon_{\rm fin} = \varepsilon_{\rm est}+\varepsilon_{\rm sm}+\varepsilon_{\rm ext}+\varepsilon_{\rm sdp}+\varepsilon_{\rm mp}.
\end{aligned} \tag{74}$$

This quantifies the certified randomness that can be securely extracted while ensuring $R_{\rm ext}$ remains private against any quantum adversary under the stated assumptions.

Bounding Eve's optimal strategy

This section establishes that the maximal single–round guessing probability attainable by an adversary in the present semi-device-independent protocol can be computed, within the stated trust model, by optimizing over quantum states and measurement operators acting on the finite Hilbert space $\mathbb{C}^3\otimes\mathbb{C}^4$. The reduction to four ancilla dimensions follows from an extremality argument applied to the complete set of linear constraints that already appeared in the previous sections, once the realistic detection process and the sequential structure of the measurements are written explicitly inside the moment matrix.

We recall that a round is accepted exactly when a single photon is found in one of the three spatial modes monitored by trusted detectors. We denote by $A_i\,(i=1,\dots,5)$ the five dichotomic observables that realise the KCBS pentagon, by A_1' the imperfectly compatible replica of A_1, and by $\eta_s\,(s=1,2,3)$ the detection efficiencies of the three signal modes.

For each accepted round, Alice chooses uniformly at random a context $X\in\{1,\dots,5\}$ from an independent trusted seed. When $X=i$, she measures the commuting pair (A_i, A_{i+1}) for $i=1,\dots,4$, and (A_5, A_1') for $i=5$, records the

raw bit $A_t = \frac{1-A_{X_t}}{2} \in \{0,1\}$, and discards the second outcome. After learning $X = i$, Eve performs a measurement described by the dichotomic observable E_i on her subsystem. Since the raw alphabet is binary, Helstrom's theorem implies that Eve never gains by employing more than two outcomes, so the operator family relevant for a single round is

$$\mathcal{M} \text{ as defined in Eq. (49).} \tag{75}$$

This refined operator set corresponds to the level-2+AB moment matrix formulation, where 'level-2' refers to monomials of degree at most 2, and '+AB' indicates that mixed monomials up to degree 2 are included to encode the intended algebra. We enforce near-commutation via (50); we do not impose exact anticommutator identities when $\epsilon_{\text{com}} > 0$. To recast all quantum-mechanical requirements into a single semidefinite optimization we work with one level-2+AB moment matrix

$$\Gamma_{XY} \;=\; \mathrm{tr}\big[\rho_{AE}^{(1)}\, X^\dagger Y\big], \tag{76}$$

built on the operator set $\mathcal{M}$ introduced above. The linear relations imposed in the code, and therefore in the proof, are stated here in compact algebraic form.

First, normalisation and idempotency are fixed by

$$\Gamma_{\mathbb{I},\mathbb{I}} = 1, \qquad A_i^2 = A_1'^2 = E_i^2 = \mathbb{I}, \tag{77}$$

Second, the experimentally determined expectation intervals are encoded as

$$\chi_{\text{KCBS}} - \varepsilon_{\text{KCBS}} \le \sum_{i=1}^{4} \Gamma_{A_i,A_{i+1}} + \Gamma_{A_5,A_1'} \le \chi_{\text{KCBS}} + \varepsilon_{\text{KCBS}}, \qquad R - \varepsilon_R \le \Gamma_{A_1,A_1'} \le R + \varepsilon_R, \tag{78}$$

where $\varepsilon_{\text{KCBS}}$ and ε_R denote the statistical half-widths of the confidence intervals for χ_{KCBS} and R, respectively. The single-body bias bounds

$$|\Gamma_{\mathbb{I},A_i}| \le \delta,\ |\Gamma_{\mathbb{I},A_1'}| \le \delta. \tag{79}$$

These relations close a convex and bounded feasible set for a 17×17 principal sub-matrix (operators: $\mathbb{I}$, $A_{1..5}$, A_1', $E_{1..5}$, and the five KCBS pair monomials), augmented by the $A_iA_{i+1}E_j$ block. Near-commutation is enforced via (50); we do not impose exact Jordan identities when $\epsilon_{\text{com}} > 0$.

Because Eve's goal is to maximise the average overlap $\mathrm{tr}\Big[(A_i \otimes E_i)\rho_{AE}^{(1)}\Big]$, the expression

$$F(\Gamma) = \sum_{i=1}^{5} \bar{w}_i\, \frac{1 + \Gamma_{A_i,E_i}}{2} \tag{80}$$

This is proportional to the single–round success probability conditioned on acceptance. At optimum, the adversarial guessing probability equals $P_{\text{guess}}^{\max} = F(\Gamma^\star)$, attained with optimizing weights $\bar{w}_i^\star$ satisfying the interval and normalization constraints. Maximising F subject to the constraints above defines a finite–dimensional semidefinite program whose variables are the entries of the principal sub-matrix of Γ supported on span $\mathcal{M}$.

At this point we prove that no larger ancilla can improve Eve's chances. Let $\Gamma = VV^\dagger$ with $V \in \mathbb{C}^{m\times r}$. If $r \ge 5$ the linear manifold generated by the constraints (77)–(79) is not full, hence there exists a non-zero Hermitian matrix H orthogonal to every constraint. For sufficiently small $\varepsilon > 0$ both perturbations $\Gamma^\pm = V(\mathbb{I} \pm \varepsilon H)V^\dagger$ remain feasible and satisfy $F(\Gamma^+) = F(\Gamma^-) = F(\Gamma)$. Therefore Γ cannot be an extreme point of the feasible set unless $r \le 4$. Consequently there is always an optimal strategy with $\rho_{AE}^{(1)}$ supported on $\mathbb{C}^3 \otimes \mathbb{C}^4$; the explicit construction uses the four basis vectors that span the range of any extremal moment matrix. The resulting SDP is solved by interior-point methods with machine precision.

Writing $P_{\text{succ}}^{\max} = F(\Gamma^\star)$ for the optimal value, the conditional min-entropy of an accepted round is bounded by Eq. (56).

The bound in Eq. (56) expresses the certified single-shot randomness rate once $P_{\text{succ}}^{\max}$ is known and is the quantity fed into the entropy-accumulation step in the next section. It remains to justify that evaluating $P_{\text{succ}}^{\max}$ can be done on a finite Hilbert space without loss.

Proposition. *Within the loss-post-selected KCBS protocol and the operator family $\mathcal{M}$ used here (including the $A_iA_{i+1}E_j$ block), an ancilla of dimension four is* sufficient *to attain the observed optima for the reported data instances.*

The SDP is implemented on the principal moment submatrix indexed by $\mathcal{M}$ (its size equals the chosen operator list). We use a commuting/tensor model: $[A_i \otimes \mathbb{I}_E,\ \mathbb{I}_A \otimes E_j] = 0$ for all i, j.

Proof. Consider the SDP in Eq. (54): maximize the affine functional $F(\Gamma) = \sum_{i=1}^{5} \bar{w}_i \, (1+\Gamma_{A_i,E_i})/2$ over the convex, closed, and bounded feasible set defined by $\Gamma \succeq 0$ and the linear constraints (50),(53). Hence an optimal solution exists; among the optimal solutions choose an extreme point $\Gamma^\star$.

Let m be the number of linearly independent affine constraints in (54). By Pataki's rank bound for SDPs, any extreme point has rank r obeying $r(r+1)/2 \leq m$. In our instance, this implies $r \leq 4$; otherwise, if $r \geq 5$, there exists a nonzero Hermitian H orthogonal to all constraints such that, for sufficiently small $\varepsilon > 0$, the perturbations $\Gamma^{\pm} = V(\mathbb{I} \pm \varepsilon H)V^\dagger$ with $\Gamma^\star = VV^\dagger$ remain feasible and satisfy $F(\Gamma^+) = F(\Gamma^-) = F(\Gamma^\star)$, contradicting extremality. Therefore $\operatorname{rank}\Gamma^\star \leq 4$.

By Gram/GNS reconstruction, any positive semidefinite moment matrix $\Gamma^\star$ of rank r admits a realization on a Hilbert space $\mathcal{H}_A \otimes \mathbb{C}^r$ with a unit vector $|\Omega\rangle$ such that $\Gamma^\star_{X,Y} = \langle\Omega|X^\dagger Y|\Omega\rangle$ for all $X, Y \in \mathcal{M}$. Since Alice's operators act on a fixed qutrit space $\mathcal{H}_A \simeq \mathbb{C}^3$, Eve's degrees of freedom are captured by the ancilla $\mathbb{C}^r$, and $r \leq 4$ shows that an ancilla of dimension four is sufficient to attain $P_{\text{succ}}^{\max}$ for the reported instances. No larger ancilla is needed to reproduce the optimum on those instances. □

[1] A. Gajda, L. Zimmermann, M. Jazayerifar, G. Winzer, H. Tian, R. Elschner, T. Richter, C. Schubert, B. Tillack, and K. Petermann, "Highly efficient CW parametric conversion at 1550 nm in SOI waveguides by reverse biased p-i-n junction," *Optics Express*, vol. 20, pp. 13100–13107, Jun 2012.

[2] E. Engin, D. Bonneau, C. M. Natarajan, A. S. Clark, M. G. Tanner, R. H. Hadfield, S. N. Dorenbos, V. Zwiller, K. Ohira, N. Suzuki, H. Yoshida, N. Iizuka, M. Ezaki, J. L. O'Brien, and M. G. Thompson, "Photon pair generation in a silicon micro-ring resonator with reverse bias enhancement," *Optics Express*, vol. 21, pp. 27826–27834, Nov 2013.

[3] J. W. Silverstone, D. Bonneau, K. Ohira, N. Suzuki, H. Yoshida, N. Iizuka, M. Ezaki, C. M. Natarajan, M. G. Tanner, R. H. Hadfield, V. Zwiller, G. D. Marshall, J. G. Rarity, J. L. O'Brien, and M. G. Thompson, "On-chip quantum interference between silicon photon-pair sources," *Nature Photonics*, vol. 8, no. 2, pp. 104–108, 2014.

[4] S. Kochen and E. P. Specker, "The problem of hidden variables in quantum mechanics," *Ernst Specker Selecta*, pp. 235–263, 1990.

[5] E. Specker, "Die logik nicht gleichzeitig entscheidbarer aussagen," *Ernst Specker Selecta*, pp. 175–182, 1990.

[6] J. Ahrens, E. Amselem, A. Cabello, and M. Bourennane, "Two fundamental experimental tests of nonclassicality with qutrits," *Scientific Reports*, vol. 3, no. 1, p. 2170, 2013.

[7] A. A. Klyachko, M. A. Can, S. Binicioğlu, and A. S. Shumovsky, "Simple test for hidden variables in spin-1 systems," *Physical Review Letter*, vol. 101, p. 020403, Jul 2008.

[8] R. Lapkiewicz, P. Li, C. Schaeff, N. K. Langford, S. Ramelow, M. Wieśniak, and A. Zeilinger, "Experimental non-classicality of an indivisible quantum system," *Nature*, vol. 474, no. 7352, pp. 490–493, 2011.

[9] D. F. V. James, P. G. Kwiat, W. J. Munro, and A. G. White, "Measurement of qubits," *Physical Review A*, vol. 64, p. 052312, Oct 2001.

[10] S. Brierley, S. Weigert, and I. Bengtsson, "All mutually unbiased bases in dimensions two to five," *arXiv preprint arXiv:0907.4097*, 2010.

[11] O. Gühne, M. Kleinmann, A. Cabello, J.-A. Larsson, G. Kirchmair, F. Zähringer, R. Gerritsma, and C. F. Roos, "Compatibility and noncontextuality for sequential measurements," *Physical Review A*, vol. 81, p. 022121, Feb 2010.

[12] S. Gupta, D. Saha, Z.-P. Xu, A. Cabello, and A. S. Majumdar, "Quantum contextuality provides communication complexity advantage," *Physical Review Letters*, vol. 130, p. 080802, 2023.

[13] M. Zahidy, D. Ribezzo, R. Müller, J. Riebesehl, A. Zavatta, M. Galili, L. K. Oxenløwe, and D. Bacco, "Single-photon-based clock analysis and recovery in quantum key distribution," *AVS Quantum Science*, vol. 5, no. 4, 2023.

[14] E. Barker, L. Bassham, A. Calis, C. Celi, and T. Hall, "Random bit generation rbg," *URL: https://csrc. nist. gov/projects/random-bitgeneration/documentation-and-software*.

[15] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.