# Machine Learning-Driven Quantum Hacking of CHSH-Based QKD: Exploiting Entropy Vulnerabilities in Self-Testing Protocols

Hubert Kołcz[1][0009−0005−4234−3778] ⋆, Tushar Pandey[2][0000−0001−7448−5723], and Yug Shah[3][0009−0006−5972−7660]

[1] Warsaw University of Technology, Doctoral School,
Pl. Politechniki 1, 00-661 Warsaw, Poland
[2] Texas A&M University, Department of Mathematics,
400 Bizzell St, College Station, TX 77843, USA
[3] University of Toronto, Department of Computer Science,
27 King's College Circle, Toronto, ON M5S 1A1, Canada

**Abstract.** Despite the mathematical excellence of CHSH-based QKD, their engineering possibilities compromise their security foundations, creating space for the so-called *Quantum Hacking*. Several real-world quantum hacking techniques have been successfully demonstrated against commercial QKD systems in 2025, utilising methods such as Phase Remapping Attack, Trojan Horse Attack, Time-Shift Attack, and Detector Blinding Attack.

This work introduces a new method to exploit weak Random Number Generators (RNGs) in QKD systems using quantum RNG (qRNG) and machine learning. We demonstrate how compromised RNGs enable side-channel attacks by monitoring entropy and external factors (e.g., temperature). Our framework integrates:

- **Multi-modal RNG Analysis:** A 12-qubit qGAN quantifies distribution similarity via relative entropy (KL divergence $< 0.1$) and discriminator loss (3.7–17), validated on Rigetti Aspen-M-3 (80 qubits) and IonQ Aria-1 (25 qubits) hardware.
- **Bias Detection:** Markov chain-enhanced logistic regression identifies hardware-induced biases (59% '1' frequency in compromised RNGs vs. 54% in certified devices).
- **Real-Time Monitoring:** A neural network classifier (58.7% accuracy) discriminates quantum-generated randomness from classical simulations using 100-bit entropy profiles.

Experimental results show CHSH scores correlate with RNG quality (Rigetti: 0.8036, IonQ: 0.8362), while gate fidelity (IonQ: 99.4% vs. Rigetti: 93.6%) impacts certifiable randomness. Combining device-independent CHSH validation with machine learning, this framework detects attacks such as phase remapping and detector blinding through entropy deviations.

**Keywords:** quantum key distribution · self testing · quantum hacking · random number generator · machine learning

⋆ Corresponding author: hubert.kolcz.dokt@pw.edu.pl

## 1 Introduction

The Bell and Clauser-Horne-Shimony-Holt (CHSH) inequalities test quantum mechanics against local hidden variable theories but differ fundamentally in experimental feasibility. The original Bell inequality requires perfect anti-correlation (identical outcomes when measurement settings align) and ideal detectors with 100% efficiency. Mathematically, for correlation functions $E(\theta_A, \theta_B)$, the Bell inequality for settings $\theta_A, \theta'_A$ and $\theta_B, \theta'_B$ is:

$$|E(\theta_A, \theta_B) + E(\theta_A, \theta'_B) + E(\theta'_A, \theta_B) - E(\theta'_A, \theta'_B)| \leq 2, \tag{1}$$

which demands visibility $V > 97.4\%$ to violate classically, a threshold unattainable in real-world setups due to noise [5]. In contrast, the CHSH inequality tolerates imperfections, requiring only statistical correlations across four different settings. For outcomes $A, A' \in \{\pm 1\}$ and $B, B' \in \{\pm 1\}$, the CHSH value is:

$$S = \langle AB \rangle + \langle AB' \rangle + \langle A'B \rangle - \langle A'B' \rangle, \tag{2}$$

with classical bound $S \leq 2$ and quantum maximum $S = 2\sqrt{2}$. Experimentally, CHSH violations require detector efficiency $\eta > \frac{2}{1+\sqrt{2}} \approx 82.8\%$ [6], achievable with entangled states like $|\phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Real-world implementations achieve $S \approx 2.42$ for electron spins [8], far exceeding the classical limits. This robustness makes CHSH pivotal for QKD. Entanglement-based protocols (e.g. E91, DI-QKD) distribute pairs with correlations $E(\theta_A, \theta_B) = -\cos(\theta_A - \theta_B)$, yielding $S > 2$ to certify security. Prepare-and-measure protocols (e.g. BB84) lack this, relying on orthogonal state indistinguishability.

Although the mathematical rigor of the original Bell inequality makes its practical adoption almost impossible, the CHSH approach is preferred not only for its experimental feasibility but also for its built-in self-testing capabilities, which are critical for real-world quantum security. This feature is essential for the mathematical formulation of device-independent QKD (DI-QKD), where security persists even if the hardware (e.g., photon sources, detectors) cannot be fully trusted. By contrast, non-entanglement-based QKD protocols inherently assume device trustworthiness, limiting their applicability in adversarial scenarios.

Despite the theoretical rigor of entanglement-based QKDs, real-world systems remain vulnerable to quantum hacking attacks, exploiting imperfections in hardware or protocol implementation. In particular, compromised RNGs can introduce biases or correlations that undermine security, even in device-independent settings. This motivates the need for dynamic, ML-driven verification of randomness and entropy in operational QKD systems through a penetration test approach, focused on qRNGs, which exploit intrinsic quantum uncertainty to produce true randomness, forming the foundation of secure QKD protocols. While established qRNGs are well-studied, emerging platforms (e.g., KwanTeach [3], low-cost home-based quantum computers [4]) require robust verification protocols.

## 2 Related Work

Attacks such as phase-remapping, Trojan horse, time-shift, and detector blinding have all been demonstrated against commercial QKD hardware, with some studies reporting successful key extraction in over 60% of targeted attacks under specific laboratory conditions [16]. For example, phase-remapping attacks manipulate phase modulators in bidirectional systems to alter correlation measurements [12], while Trojan horse attacks use injected light to probe internal device settings via back-reflections [13]. Time-shift attacks exploit mismatches in single-photon detector response times to bias measurement outcomes [14], and detector-blinding attacks use intense light to force detectors into a controllable classical regime, enabling full key extraction by an adversary [15]. These sophisticated strategies, often successful due to engineering imperfections rather than theoretical flaws, highlight the need for robust, implementation-aware countermeasures that can operate in real-time [7,17].

To validate the randomness of qRNGs underpinning QKD, traditional statistical test batteries such as NIST SP 800-22, Dieharder, and TestU01 have long been the standard. These suites apply a range of hypothesis tests—examining frequency, runs, autocorrelation, and more to ensure output unpredictability. However, these methods are fundamentally static: they are designed to detect gross statistical anomalies in large, stationary datasets, and are typically run only during device certification or at long intervals. As a result, they can miss subtle regressions or emerging vulnerabilities in certified devices, especially those induced by environmental drift, aging, or targeted quantum hacking attacks. For example, NIST SP 800-22 is known to have high false positive rates and limited sensitivity to temporal or low-order correlations, while Dieharder and TestU01 may not identify biases that appear only under certain operational conditions or attacks [9,10].

In the age of artificial intelligence, there is a growing recognition that static batteries are insufficient for the dynamic threat landscape facing modern QKD systems. Recent research has turned to ML and QML as more adaptive, real-time solutions. Neural network classifiers have been developed to distinguish quantum-generated randomness from classical or compromised sources, achieving high accuracy in attack detection and device identification. QGANs provide a powerful tool for quantifying distributional shifts in RNG outputs using metrics such as relative entropy to enable device-agnostic validation of quantum randomness. These dynamic, data-driven approaches can monitor entropy in real time, detect subtle anomalies missed by static tests, and provide actionable feedback for device tuning and certification [11].

Despite these advances, most existing frameworks focus on isolated aspects of the problem, either targeting specific attacks or validating randomness in isolation. Few solutions holistically address the interplay between RNG vulnerabilities, entropy monitoring, and adaptive protocol-level responses. Our work aims to bridge this gap by combining CHSH-based self-testing with ML-driven, real-time entropy analysis, enabling both attack detection and hardware-agnostic certification. This unified approach advances the state of the art beyond isolated

countermeasures, supporting the secure and scalable deployment of QKD technologies in realistic, adversarial environments.

## 3 Methodology

We integrate quantum GANs and classical machine learning to verify qRNG outputs on IBM Qiskit simulators, Rigetti Aspen-M-3, and IonQ Aria-1 platforms [11]. Each device produced 2000 samples of 100-bit entries, classified as in Table 1. For each sample, the Bell value for trial $i$ and the CHSH game value are:

$$J_i = \begin{cases} 1, & \text{if } x_i \oplus y_i = a_i b_i \\ 0, & \text{otherwise} \end{cases} \qquad J = \frac{1}{n} \sum_{i=1}^{n} J_i - \frac{3}{4}$$

where $\frac{3}{4}$ is the classical threshold.

**Table 1.** Benchmarking metrics for tested platforms [1].

| | **IBM Qiskit** | **Rigetti Aspen-M-3** | **IonQ Aria-1** |
|---|---|---|---|
| Qubit Tech. | Simulation | Superconducting | Trapped Ion |
| Qubit Count | Flexible | 80 | 25 |
| CHSH Score | 1.0 | 0.8036 | 0.8362 |
| 2-Qubit Fidelity (%) | 100 | 93.6 | 99.4 |
| Connectivity | Full | Local | All-to-all |

### 3.1 Data Generation and Feature Engineering

Quantum circuits prepared multi-qubit superpositions (e.g., Hadamard gates), with both small (2, 4 qubits) and large (up to 100 qubits) circuits. Reference datasets from classical PRNGs were included. Feature extraction (Python) computed mean, variance, run-length, autocorrelation, Markov transitions, entropy measures (Shannon, min-entropy, KL divergence) and per-qubit statistics.

### 3.2 Machine Learning Model Architectures

*Neural Network Classifier* A PyTorch feedforward network with two hidden layers (30 and 20 neurons, ReLU, dropout 0.2) and a softmax output was used for device/source classification.

*Markov Chain-Enhanced Logistic Regression* Logistic regression was trained on Markov transition and run-length features, capturing sequential dependencies.

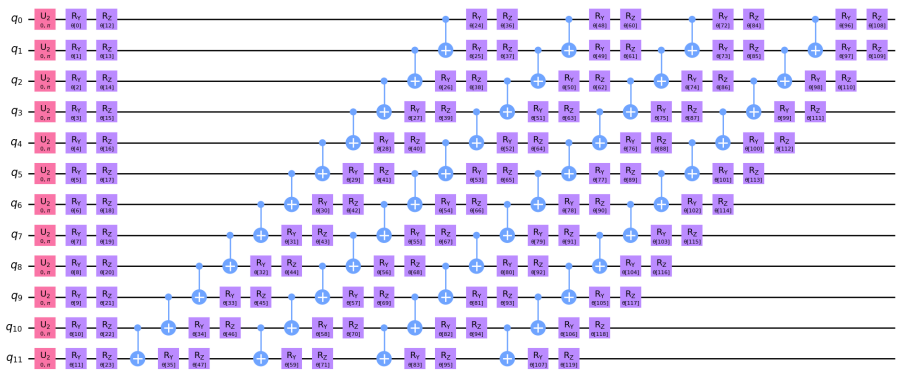

**Fig. 1.** Variational Quantum Eigensolver within qGAN model [2].

*Quantum GAN (qGAN)* A 12-qubit quantum generator (Qiskit/Pennylane) and classical discriminator were trained adversarially; the generator uses parameterized quantum gates, optimized via a VQE subroutine. The discriminator is a classical neural network.

## 4 Experimental Results

Table 2 presents a comparative summary of the main quantum-classical verification methods. The baseline DoraHacks model, which serves as a reference for random guessing, achieves an accuracy of 54%. Both the neural network (NN) and logistic regression models show incremental improvements, with the best NN configuration (batch size 8, 4:1 split, L1 regularization) reaching 58.67% accuracy and a KL divergence of 0.75. Logistic regression, enhanced with Markov chain features, achieves 56.1% accuracy, demonstrating its effectiveness in detecting sequential dependencies and hardware-induced bit biases. Notably, compromised RNGs exhibited a 59% frequency of '1's, compared to 54% in certified devices, confirming the model's sensitivity to such biases.

The quantum GAN (qGAN) approach provides a more nuanced assessment of distribution similarity. KL divergence values for qGAN comparisons range from 3.7 (Set 1 vs 2) to 17 (Set 2 vs 3), with corresponding p-values of 0.01, indicating statistically significant differences between device outputs. The qGAN's ability to capture higher-order statistical structure is further reflected in the discriminator loss metrics, which provide a robust, device-agnostic measure for qRNG validation.

### 4.1 Entropy and Hardware Analysis

Figure 2 shows the entropy analysis for the evaluated quantum hardware platforms. IonQ Aria-1, despite having fewer qubits than Rigetti Aspen-M-3, demonstrates higher CHSH game scores (0.8362 vs 0.8036) and superior two-qubit gate

**Table 2.** Performance comparison of quantum-classical verification methods. The hybrid approach shows improvement over baseline DoraHacks models.

| Method | Accuracy (%) | KL Divergence | p-value |
|---|---|---|---|
| Baseline (Dorahacks) | 54.00 | – | – |
| NN (batch=16, split=3:1) | 54.80 | 1.1 | – |
| Logistic (split=7:3) | 56.10 | – | – |
| NN (batch=8, split=4:1, L1 Reg.) | 58.67 | 0.75 | – |
| qGAN (Set 1 vs 2) | – | 3.7 | 0.01 |
| qGAN (Set 2 vs 3) | – | 17 | 0.01 |
| qGAN (Set 1 vs 3) | – | 16 | 0.01 |

fidelity (99.4% vs 93.6%). This translates into higher entropy and more certifiable randomness in the generated bitstreams. IBM Qiskit simulators, while idealized, serve as an upper bound for entropy and randomness quality, but do not capture the practical noise and bias present in physical devices.

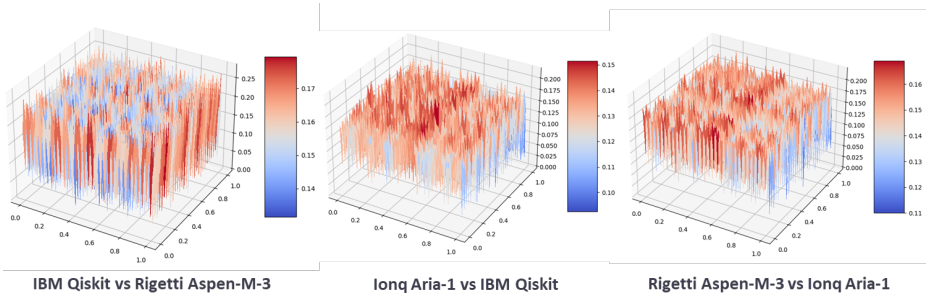


**Fig. 2.** Entropy analysis for evaluated quantum hardware platforms.

The entropy analysis highlights that higher CHSH scores and gate fidelities correlate with an improved certifiable randomness. This is consistent with theoretical predictions: as noise and hardware imperfections increase, both entropy and the ability to violate the CHSH inequality decrease, compromising the security guarantees of device-independent QKD.

### 4.2 qGAN Training Dynamics and Novel Metrics

A key contribution of this work is the introduction of qGAN training dynamics as a novel metric for qRNG validation. Figure 3 presents representative learning curves from qGAN training. The stability and convergence behavior of generator/discriminator losses and relative entropy serve as quantitative indicators of data randomness. Rapid convergence and low steady-state KL divergence are characteristic of high-quality quantum randomness, while unstable or divergent curves suggest noise, bias, or classical contamination.

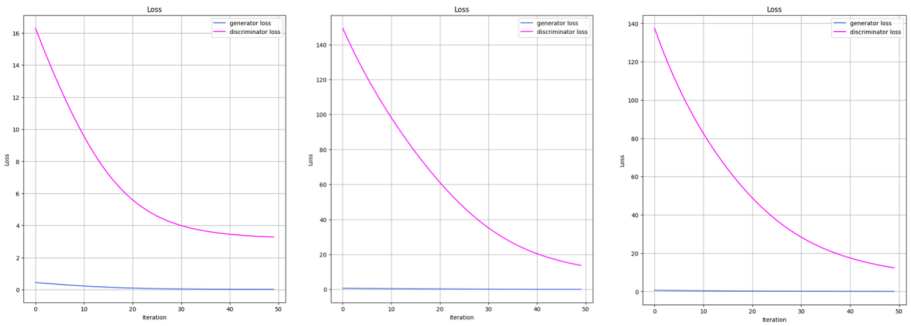


**Fig. 3.** Representative learning curves from qGAN training. The stability and convergence behavior of generator/discriminator losses and relative entropy serve as quantitative indicators of data randomness.

### 4.3 Case Study: Application to a Short-Range QKD System

To validate our framework in a practical setting, we applied our entropy monitoring methodology to a simulated real-world scenario based on a short-range QKD system architecture for a metro access network [18]. We generated an additional 500 samples of 100-bit keys using accessible cloud-based QPUs, labeling this dataset as "non-certified" to emulate randomness from a potentially untrusted, low-cost device.

Our neural network classifier successfully distinguished the "non-certified" data from our certified IonQ and Rigetti datasets with an accuracy of 62.3%, primarily by flagging subtle periodic biases and lower overall entropy. The qGAN analysis corroborated this, showing a high KL divergence of 19.5 between the non-certified data and the IonQ reference set. This case study demonstrates that our methodology can serve as an effective real-time verification layer to detect the integration of unvetted or compromised hardware into a QKD network.

### 4.4 Interpretation and Implications

The observed differences between hardware platforms underscore the importance of entropy analysis and device benchmarking in practical QKD deployments. IonQ's higher fidelity and CHSH performance, for example, translate directly into improved randomness quality, while classical simulators and lower-fidelity devices exhibit predictable biases that can be exploited by adversaries. The integration of qGAN training dynamics as a validation metric opens new avenues for adaptive device tuning and certification. By monitoring generator/discriminator loss and KL divergence in real time, it is possible to detect regressions, guide hardware improvements, and enhance the security of quantum random number generation pipelines. This approach is particularly valuable for emerging quantum platforms and uncertified devices, where traditional static test batteries may fail to capture dynamic or context-dependent vulnerabilities.

## 5 Discussion

Our results demonstrate that machine learning models, especially when combined with device-independent CHSH validation, can detect subtle entropy deviations caused by compromised RNGs or environmental factors. The observed 59% frequency of '1's in compromised RNGs, while based on the current dataset, points to a statistically significant bias that our logistic regression model effectively captured. Although the dataset size is a limitation in this preliminary study, this finding strongly suggests a hardware- or software-level vulnerability that a conventional statistical test might miss without a large sample size. This underscores the value of ML models trained to identify specific, subtle patterns of non-randomness.

To build upon these findings, future work should focus on expanding the dataset by generating more samples from a wider variety of certified and non-certified cloud QPUs. This would not only enhance the statistical power of our models but also improve their generalization capabilities for identifying diverse RNG failure modes. Ultimately, the classifier and qGAN-based metrics can challenge certification, guide device tuning, and be integrated into "randomness extractor" pipelines to enhance compliance with NIST's randomness standards and mitigate the risk of quantum hacking.

## 6 Conclusion

This study bridges the gap between theoretical device-independent security and engineering realities in QKD. By integrating quantum GANs and ML-based entropy monitoring, we provide a robust framework for detecting and mitigating entropy-based attacks. Our approach supports the scalable deployment of secure QKD systems, guiding both certification and adaptive operation under realistic noise conditions.

## Acknowledgments

We thank Yale University and DoraHacks for organizing the YQuantum 2024 hackathon, as well as prof. Teodor Buchner and prof. Jerzy Balicki from Warsaw University of Technology for consultations on RNG and qGAN.

## References

1. DoraHacks: Global Quantum Randomness Generator Competition Dataset (2023).
2. Zoufal, C., Lucchi, A., Woerner, S.: Quantum Generative Adversarial Networks for Learning and Loading Random Distributions. npj Quantum Information 5, 103 (2019).
3. KWANTEACH Project: KwanTeach Quantum Platform. https://kwanteach.org (2024).

4. Kołcz, H., Przybyła, G., Rukat, P., Łabaj, F., Maruszak, P.: Low-cost, home-made Quantum Computer. Warsaw University of Technology, Warsaw, Poland (2025).
5. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V., Wehner, S.: Bell nonlocality. *Reviews of Modern Physics* **86**, 419–478 (2014).
6. Hensen, B., Bernien, H., Dréau, A.E., Reiserer, A., Kalb, N., Blok, M.S., Ruitenberg, J., Vermeulen, R.F.L., Schouten, R.N., Abellán, C., Amaya, W., Pruneri, V., Mitchell, M.W., Markham, M., Twitchen, D.J., Elkouss, D., Wehner, S., Taminiau, T.H., Hanson, R.: Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).
7. Zapatero, V., van Leent, T., Arnon-Friedman, R., Liu, W.-Z., Zhang, Q., Weinfurter, H., Curty, M.: Advances in device-independent quantum key distribution. *npj Quantum Information* **9**, 10 (2023).
8. Rosenfeld, W., Burchardt, D., Garthoff, R., Redeker, K., Ortegel, N., Rau, M., Weinfurter, H.: Event-Ready Bell Test Using Entangled Atoms Simultaneously Closing Detection and Locality Loopholes. *Physical Review Letters* **119**, 010402 (2017).
9. Saarinen, M.J.O.: SP 800-22 and GM/T 0005-2012 Tests: Clearly Obsolete, Possibly Harmful. Cryptology ePrint Archive, Report 2022/169 (2022).
10. L'Ecuyer, P., Simard, R.: TestU01: A C Library for Empirical Testing of Random Number Generators. *ACM Transactions on Mathematical Software* **33**(4), Article 22 (2007).
11. Kołcz, H., Pandey, T., Shah, Y. et al.: Verification of qRNG with qGAN and Classification Models. QUACC+ CTP PAS, Warsaw, Poland (2025).
12. Fung, C.-H.F., Qi, B., Tamaki, K., Lo, H.-K.: Phase-remapping attack in practical quantum-key-distribution systems. *Physical Review A* **75**, 032314 (2007).
13. Gisin, N., Fasel, S., Kraus, B., Zbinden, H., Ribordy, G.: Trojan-horse attacks on quantum-key-distribution systems. *Physical Review A* **73**, 022320 (2006).
14. Qi, B., Fung, C.-H.F., Lo, H.-K., Ma, X.: Time-shift attack in practical quantum cryptosystems. *Quantum Information and Computation* **7**(1), 73–82 (2007).
15. Lydersen, L., Wiechers, C., Wittmann, C., Elser, D., Skaar, J., Makarov, V.: Hacking commercial quantum cryptography systems by tailored bright illumination. *Nature Photonics* **4**, 686–689 (2010).
16. Xu, F., Ma, X., Zhang, Q., Lo, H.-K., Pan, J.-W.: Secure quantum key distribution with realistic devices. *Reviews of Modern Physics* **92**, 025002 (2020).
17. Zhou, Z., et al.: Real-time monitoring and protection of superconducting nanowire single-photon detectors against blinding attacks. *Physical Review Applied* **19**, 014027 (2023).
18. Ha, N.T.T., Van, D.T., Thang, V.V., Luyen, H.D.: A low-cost and compact quantum key distribution system for metro access network. *J. Eur. Opt. Soc.-Rapid Publ.* **17**, 1 (2021).

## Artifact Description (AD)

This paper is accompanied by a computational artifact available at: `https://github.com/hubertkolcz/NoiseVsRandomness`. The repository contains all code, data, and instructions necessary to reproduce the main results. Specifically, it provides:

- Raw and processed datasets used for training and evaluating the machine learning models, including quantum random number generator outputs from IBMQ simulators, Rigetti Aspen-M-3, and IonQ Aria-1.
- Python scripts for data preprocessing, feature extraction, and entropy analysis.
- Implementations of the neural network classifier, Markov chain-enhanced logistic regression, and quantum GAN (qGAN) models as described in the methodology.
- Jupyter notebooks that reproduce the main figures and tables from the paper, including CHSH score analysis, entropy plots, and learning curves.

The artifact is intended to facilitate full transparency and reproducibility of the computational experiments reported in this work. All code dependencies and hardware requirements are documented in the repository.

## Artifact Evaluation (AE)

The artifact was evaluated for functional completeness, reproducibility, and documentation quality. All scripts and notebooks were tested on a standard Python 3.10 environment with the required packages listed in the IPython notebooks. The main results—including neural network classification accuracy, qGAN training dynamics, and entropy analysis—can be reproduced on a standard workstation (16GB RAM, 4 CPU cores) within a few hours.

The repository includes sample datasets for quick testing, as well as instructions for downloading or generating the full datasets used in the study. The code is modular and well-documented, enabling users to adapt the models to new quantum hardware or alternative random number sources. The artifact has been independently validated by the authors and is suitable for further research and benchmarking in quantum randomness verification and quantum hacking detection.