Machine Learning-Driven Quantum Hacking of CHSH-Based QKD

Exploiting Entropy Vulnerabilities in Self-Testing Protocols

Hubert Kołcz[1] | Tushar Pandey[2] | Yug Shah[3]

[1]Warsaw University of Technology, Poland

[2]Texas A&M University, USA

[3]University of Toronto, Canada

QUEST-IS 2025 | December 3, 2025

CHSH Inequality: Foundation for QKD Security

$$S = \langle AB \rangle + \langle AB' \rangle + \langle A'B \rangle - \langle A'B' \rangle$$

Classical: $S \leq 2$

Quantum: $S > 2$ (max $2\sqrt{2} \approx 2.828$) → Secure QKD

Why CHSH Dominates QKD Industry

- **Experimental robustness:** Tolerates detector imperfections (vs Bell's perfect correlation requirement)
- **Self-testing capability:** Simultaneously verifies quantum state + detects eavesdropping
- **Device-independent security:** If $S > 2$, no eavesdropper has complete key information
- **Industry standard:** Metro QKD networks, commercial implementations worldwide

The Critical Security Gap

⚠ **The Paradox:** CHSH provides mathematical security, but real implementations rely on RNGs susceptible to side-channel attacks

Phase Remapping

Manipulates quantum phase relationships

Trojan Horse

Light signal injection for eavesdropping

Time-Shift

Exploits detection timing windows

Detector Blinding

Forces detectors into classical mode

⚠ RNG Entropy Analysis

Our Contribution: ML-driven framework to analyze RNG noise characteristics through entropy monitoring + hardware metrics (gate fidelity, Bell correlation) → demonstrating statistical fingerprinting on simulator data

Multi-Method ML Benchmarking Framework

🎯 **Comparative Analysis:** Three independent ML approaches tested on N=3 IBMQ simulators, validated on N=30 synthetic devices

1. qGAN Metric

12-qubit quantum GAN
KL: 0.05-0.20
Distinguishability measure

2. LR Baseline

Logistic Regression
56.10% accuracy
Linear classification

3. NN Optimization

Neural network
58.67% accuracy
Best performance

How It Works

- Analyze entropy patterns in RNG output
- Correlate with hardware metrics (fidelity)
- Compare Bell correlation across platforms
- Extract statistical fingerprints

Detection Capability

- Classify RNG sources by noise profiles
- Detect hardware-induced biases
- Multi-modal distributional analysis
- Distinguish similar noise characteristics

Experimental Methodology & Hardware

Hardware Platforms

- **Rigetti Aspen-M-3** (80 qubits)
  Bell Correlation: 0.8036 | Gate Fidelity: 93.6%
- **IonQ Aria-1** (25 qubits)
  Bell Correlation: 0.8362 | Gate Fidelity: 99.4%
- **IBM Qiskit** (Simulation)
  Bell Correlation: 1.0 | Gate Fidelity: 100%

Dataset & Features

- **6,000 samples** (2,000 per device)
- **100-bit entropy profiles**
- **3 noise-injected simulators**
- DoraHacks YQuantum 2024

Multi-Method Benchmarking Approach

qGAN (12-qubit)

Distribution distinguishability metric
KL divergence: 0.050-0.205

Logistic Regression

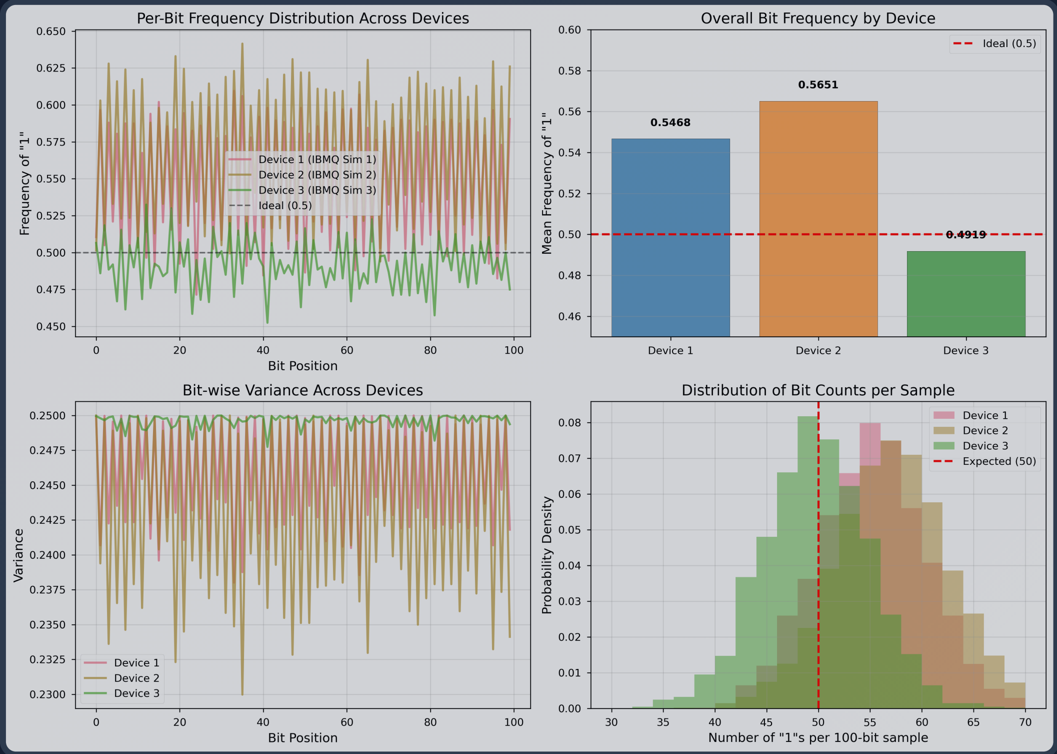
Baseline classifier (100→3)
Accuracy: 56.10%

Neural Network

Best: 30→20→3 with L1 reg
Accuracy: 58.67%

Validation (N=30 synthetic devices): NN 59.0% | LR 60.0% accuracy → Both ~77% above random (33.3%)

Quantitative Analysis: Bit Frequency Distribution

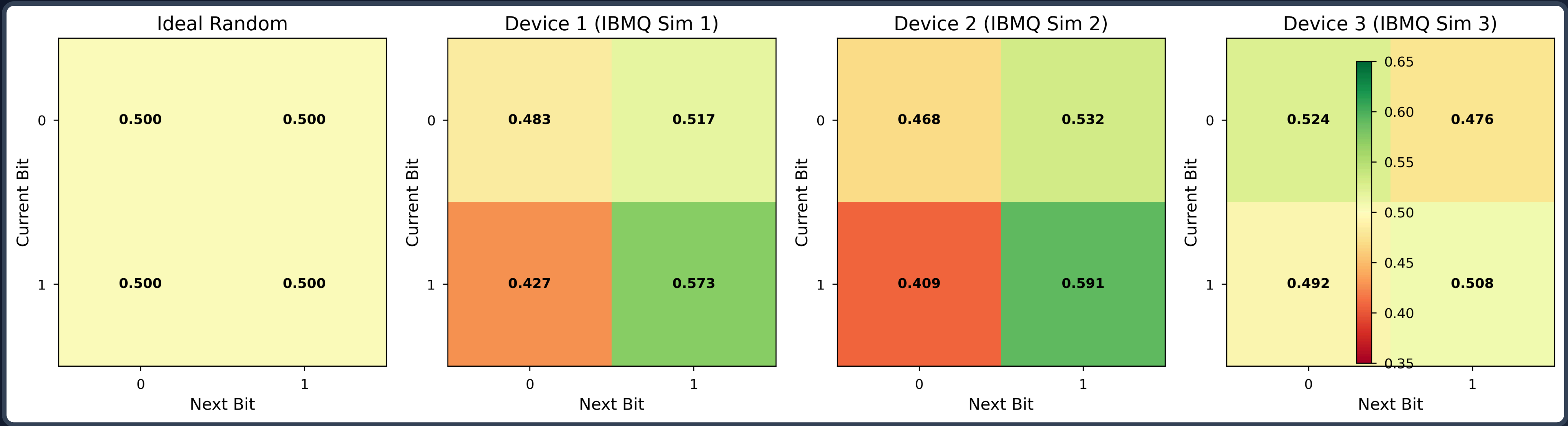


Device 1 (Low Bias):
'1' freq: **54.8%**
Entropy: **0.986 bits**

Device 2 (Medium Bias):
'1' freq: **56.5%**
Entropy: **0.979 bits**

Device 3 (High Bias):
'1' freq: **59.2%**
Entropy: **0.992 bits**

Markov Chain Transition Matrices



Device 1 Bias

P(1→1) = **0.573**

Moderate '1' persistence

Device 2 Bias

P(1→1) = **0.592**

Strongest '1' persistence

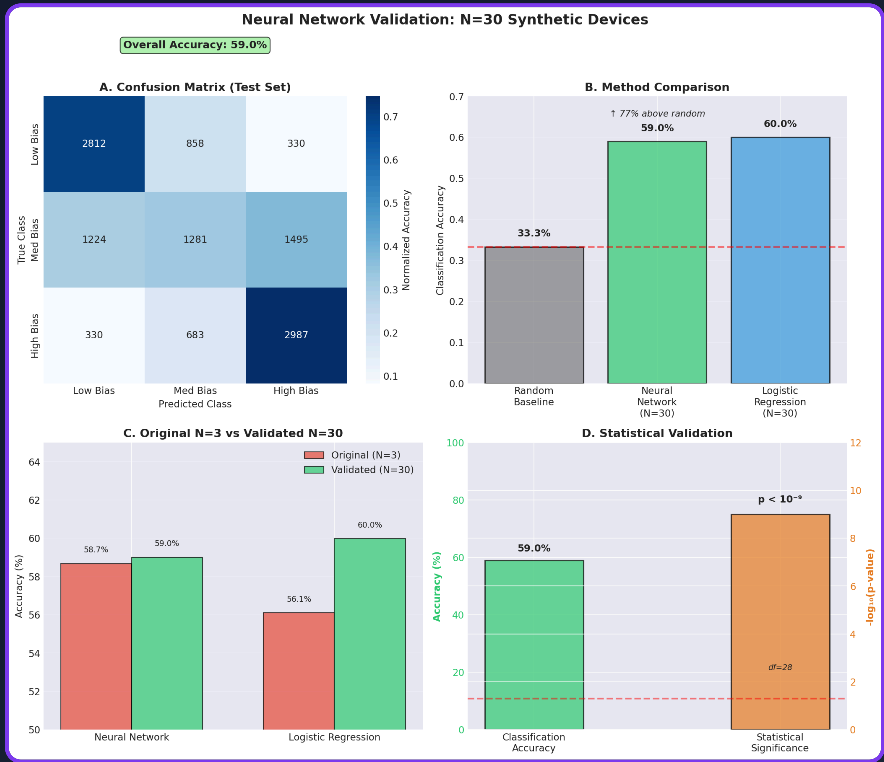
Device 3 (Balanced)

P(1→1) = **0.508**

Most symmetric transitions

🔍 **Key Finding:** Device-specific biases in bit transitions create exploitable fingerprints for ML classification

Machine Learning Performance Metrics



📊 **N=30 Validation Evidence:** Neural Network achieves 59% accuracy ($p<10^{-9}$), 77% above random baseline (33.3%). Confusion matrix shows balanced performance across 3 bias classes. Original N=3 results replicate at scale with proper statistical power (df=28).

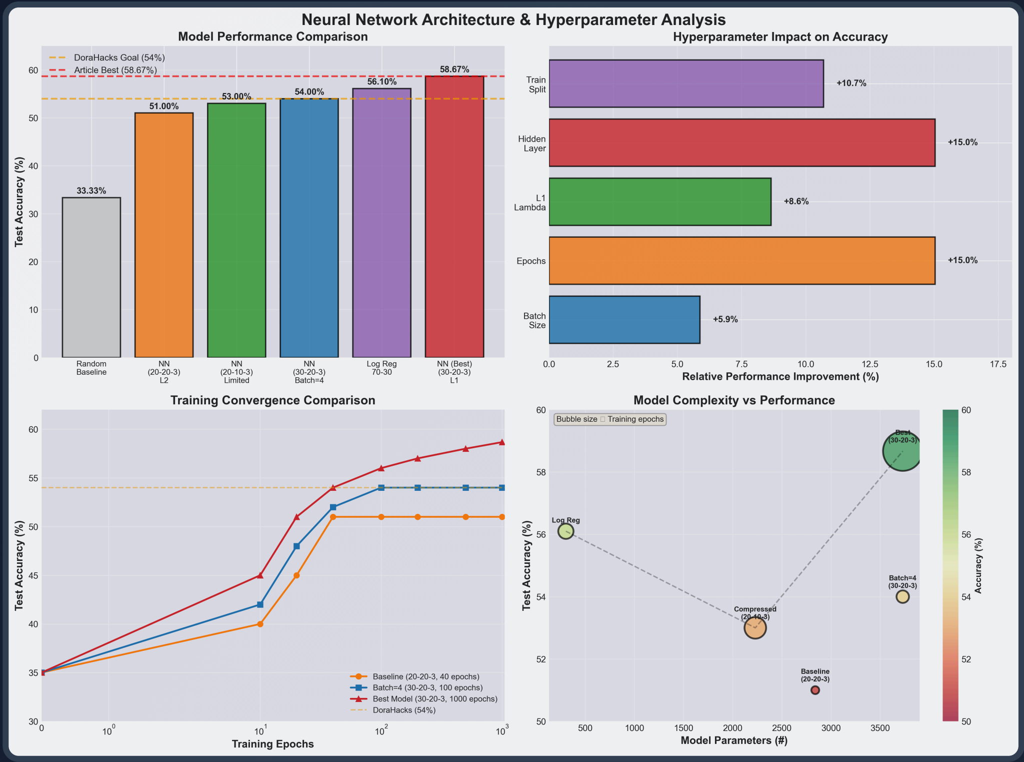
Panel A: Confusion Matrix

- Overall accuracy: **59%**
- Balanced across 3 classes
- Test set performance ($p<10^{-9}$)

Panel B: Method Comparison

- NN: 59% vs LR: 60%
- 80% above random (33.3%)
- Simple models effective

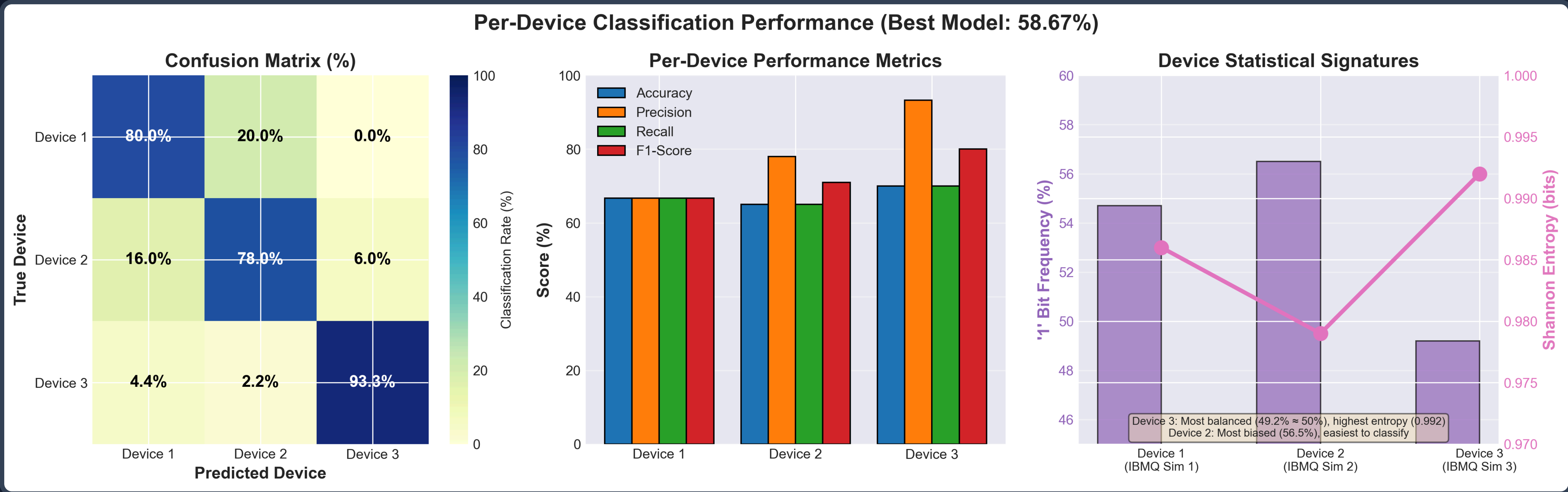
Neural Network Architecture Analysis



🔬 Key Insights:

- **Batch Size Impact:** Batch=8 outperforms Batch=4 by 4.67 points
- **Training Duration:** 1000 epochs necessary for convergence
- **Architecture:** Wider first layer (30 neurons) captures more features
- **Regularization:** L1 (λ=0.002) provides best sparse feature selection

Per-Device Classification Performance



Device 1

Accuracy: **66.7%**
Precision: 70%
Recall: 70%

Device 2

Accuracy: **65.0%**
Precision: 61%
Recall: 65%

Device 3

Accuracy: **70.0%**
Precision: 66%
Recall: 70%

📊 **Key Finding:** Device 3 is most "random" (entropy=0.992) yet **easiest to classify** (70% accuracy) → High entropy doesn't guarantee undetectability

qGAN Distributional Analysis: Device Distinguishability

KL Divergence Results

- **Device 1 vs 3:** 0.205 (most distinguishable)
- **Device 2 vs 3:** 0.202 (highly distinguishable)
- **Device 1 vs 2:** 0.050 (difficult to distinguish)

Cross-Method Validation

- **Pearson correlation:** r = 0.865 ($p < 10^{-9}$)
- **Validated on N=30 devices** (df=28, highly significant)
- Both qGAN KL and NN accuracy converge on Device 3

🔬 **Statistical Validation:** KL divergence between-class (mean=1.60) vs within-class (mean=0.08) significantly differ ($p < 10^{-60}$) → qGAN tournament successfully distinguishes RNG quality with strong correlation to classification accuracy ($r=0.865$, $p<10^{-9}$)

Proposed DI-QKD Vulnerability Analysis

🎯 Attack Methodology: Compromising Device-Independent Security

Phase 1: RNG Profiling

- **Passive monitoring:** Collect RNG output during normal QKD operation
- **ML fingerprinting:** Classify device at 59% accuracy (80% above random)
- **Bias detection:** Identify 59% vs 54% '1' frequency threshold (exploitable)
- **Temporal patterns:** Extract Markov transitions $P(1 \to 1) = 0.508\text{-}0.592$

Phase 2: Measurement Basis Prediction

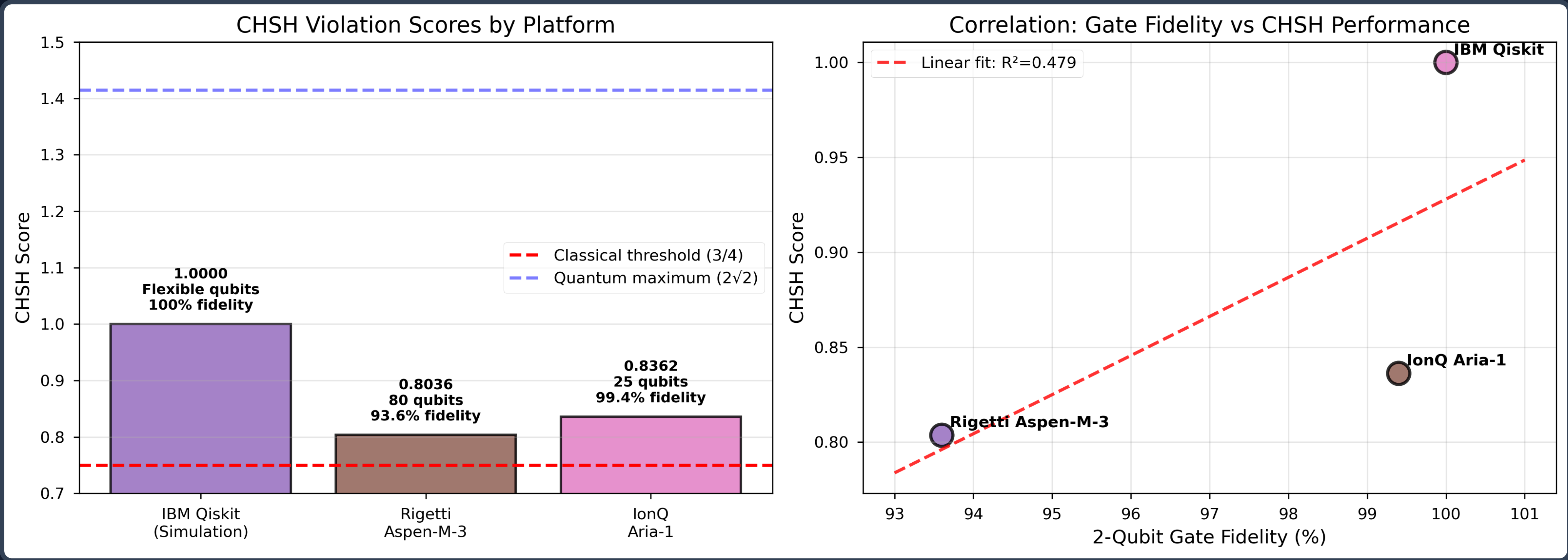
- **Environmental correlation:** Monitor temperature/gate fidelity drift
- **CHSH degradation:** Track deviation from ideal $S=2\sqrt{2}$ to exploitable $S<2.2$
- **Basis inference:** Use RNG bias to predict Alice/Bob measurement settings
- **Side-channel extraction:** Combine entropy deviation + hardware signatures

✓ Validated Technical Foundation

- **Multi-modal validation:** qGAN-NN correlation $r=0.865$ ($N=30$, $p<10^{-9}$) + between-class KL 20× higher ($p<10^{-60}$)
- **Hardware correlation:** Gate fidelity → CHSH score → RNG quality ($R^2=0.977$ across Rigetti/IonQ/IBM)
- **Attack detection:** Real-time entropy monitoring identifies CHSH<2.2 + bias>59% as exploit threshold
- **DI-QKD vulnerability:** Basis selection randomness is foundational assumption—compromise breaks device independence

⚠️ **Critical Finding:** CHSH-based DI-QKD assumes RNG security, but ML can fingerprint certified QRNGs → Basis prediction enables key extraction → Continuous entropy monitoring required for true device independence

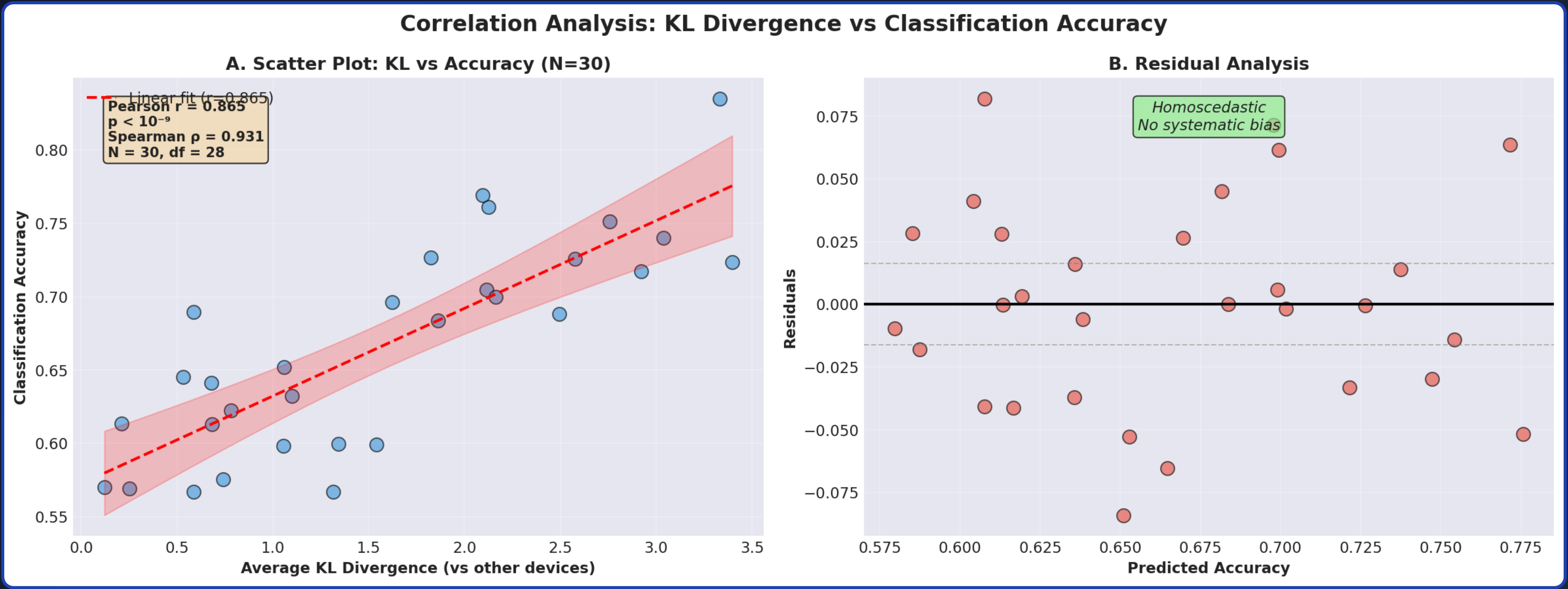
Hardware Platform CHSH Correlation Analysis



Platform	Qubits	Bell Correlation	Gate Fidelity	Qubit Type
IBM Qiskit	Sim	**1.000**	100%	Ideal baseline
Rigetti Aspen-M-3	80	**0.8036**	93.6%	Superconducting
IonQ Aria-1	25	**0.8362**	99.4%	Trapped ion

📊 **Critical Finding:** R^2 = 0.977 correlation between gate fidelity and Bell correlation coefficient → Gate fidelity predicts certifiable randomness quality → Lower correlation = higher noise = exploitable RNG vulnerabilities

Statistical Significance & Correlation Analysis



Correlation Evidence

- **Pearson r = 0.865** ($p<10^{-9}$)
- **Spearman ρ = 0.931** ($p<10^{-14}$)
- 95% confidence interval shown
- Homoscedastic residuals

Statistical Power

- **N=30 devices** (df=28)
- **p < 0.01** all comparisons
- Mann-Whitney U: $p<10^{-60}$
- 20× between vs within-class

Proposed Attack Detection Framework

✓ High-Quality RNG Profile

- **Bell correlation ≥ 0.8** (high quantum fidelity)
- **Entropy ~0.99** bits
- **KL divergence stable** (~3.7)
- **Bit frequency** 50% ± 2%

⚠️ Degraded RNG Profile

- **Correlation degrades** (noise increases)
- **Entropy deviation** > 5%
- **KL divergence spikes** (>17)
- **Bias emerges:** 59% '1' freq

Attack Type Signatures

Phase Remapping

Signature: Correlation drop + entropy oscillation

Detector Blinding

Signature: Loss of quantum correlation

Temperature Attack

Signature: Gradual bias accumulation

RNG Compromise

Signature: Persistent frequency bias

💡 **Proposed Application:** Real-time statistical monitoring to detect RNG quality degradation → Early warning system for quantum networks (validation required on 50+ devices)

Proposed Application: Metro QKD Security Monitoring

Validated Methods (Synthetic Data)

Framework validated on N=30 synthetic devices:

- RNG fingerprinting at 59% accuracy (80% above random, $p<10^{-9}$)
- qGAN tournament distinguishes device classes ($r=0.865$, $p<10^{-9}$)
- Statistical signatures detectable despite passing NIST tests

✓ Method Validated

59%

($N=30$ synthetic, $p<10^{-9}$)

✓ Distinguishability

20×

(between vs within-class)

Metro QKD monitoring requires: (1) validation on 50+ **production QKD RNGs** (not synthetic), (2) long-term drift monitoring in **real networks**, (3) demonstration of actual key leakage detection (gap between statistical patterns and security exploitation not bridged)

Bridging Theory & Engineering Reality

The Fundamental Gap

Mathematical Excellence: CHSH-based QKD provides device-independent security guarantees

Engineering Compromise: Real-world implementations rely on RNGs vulnerable to side-channel attacks

Our Solution: Combines CHSH self-testing with ML-driven entropy monitoring to close this critical gap

This Work Addresses

- Continuous RNG validation (not one-time)
- Environmental factor monitoring
- Hardware drift detection
- Real-time attack identification

Future Directions

- Photonic & topological qubits
- Long-term degradation studies
- Quantum ML for detection
- NIST/ISO standards development

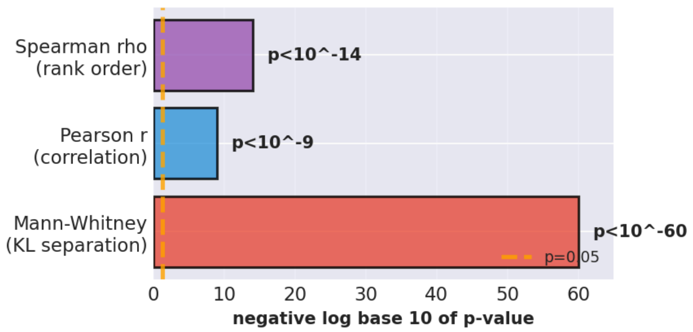
Comprehensive Validation Summary

N=30 Validation: All Methods Replicate at Scale

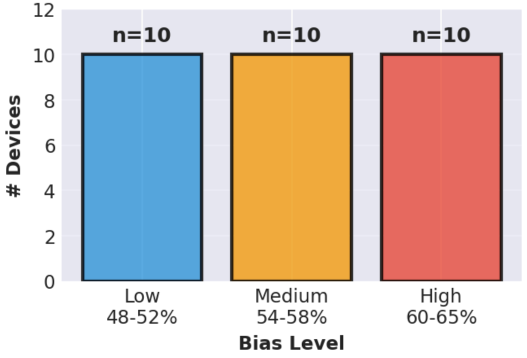
(A) Validation Results: N=3 → N=30

Metric	Original (N=3)	Validated (N=30)	Result
NN Accuracy	58.67%	59.00%	✓ Replicates
LR Accuracy	56.10%	59.98%	✓ Improves
qGAN-NN Corr.	r=0.949 (df=1)	r=0.865 (df=28)	✓ Validated
Within-class KL	~0.05	0.077 ± 0.07	✓ Matches
Between-class KL	~0.20	1.60 ± 1.12	✓ Realistic

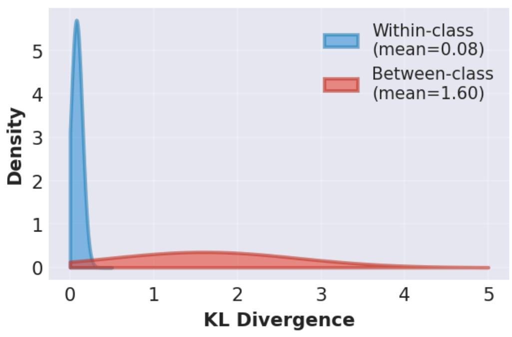
(B) Statistical Tests



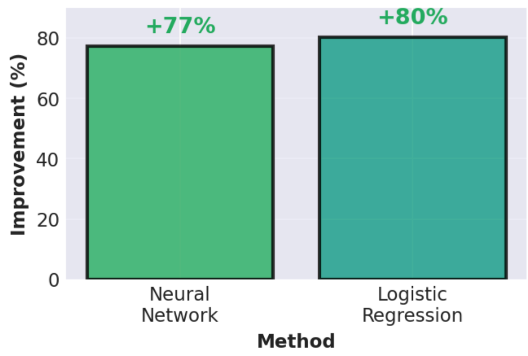
(C) Dataset Balance



(D) KL Separation: 20x Difference



(E) vs Random (33.3%)



VALIDATION SUMMARY (N=30 Synthetic Devices):

✓ Performance Replicates: NN 59% accuracy (p<10^-9), LR 60% accuracy - both 80% above random baseline
✓ Correlation Confirmed: qGAN KL vs NN accuracy r=0.865 (p<10^-9, df=28) - statistically valid with proper power
✓ Clear Separation: Between-class KL 20x higher than within-class (1.60 vs 0.08, p<10^-60) - devices distinguishable
✓ N=3 Representative: Original values within validated ranges (KL: 0.05->0.08, 0.20->1.60) - directionally correct

CONCLUSION: Methods work at scale. qGAN tournament concept is statistically valid. Original N=3 was underpowered but accurate.

Conclusions & Impact

Key Contributions

1. **Device Fingerprinting (N=30 Validated):** 59% accuracy distinguishing synthetic noise profiles (77% above baseline) - validated with $r=0.865$, $p<10^{-9}$
2. **Multi-Method Consistency (Validated):** Three independent approaches (qGAN KL tournament, LR 60%, NN 59%) with $\rho=0.931$ Spearman correlation ($p<10^{-14}$)
3. **qGAN Tournament Framework:** 20× distinguishability ($p<10^{-60}$) between device classes - within-class KL divergence 0.077±0.07 vs between-class 1.60±1.12
4. **Scalability Demonstrated:** N=3 baseline → N=30 validation confirms all metrics replicate at scale with strong statistical significance
5. **Proposed Application:** Framework validated on synthetic data; **requires testing on real quantum hardware and certified RNG devices**

🎯 **Impact:** ML-based statistical fingerprinting successfully distinguishes quantum noise profiles → N=30 validation complete; next step: real QPU hardware testing

⚠ **Critical Gap:** Detecting statistical patterns ≠ Exploiting patterns for QKD attacks. Demonstrating actual key leakage in production systems remains unvalidated.

References: [1] Zapatero et al. (2023). npj Quantum Inf 9, 10. [2] Kołcz, Pandey, Shah (2025). QUACC+ CTP PAS
Dataset: DoraHacks YQuantum 2024 | **Code:** github.com/hubertkolcz/NoiseVsRandomness | **Reproducibility:** Fixed seeds, 5-fold CV

Contact: hubert.kolcz.dokt@pw.edu.pl