

# **ML-Driven Quantum Hacking of CHSH-Based QKD**

Speaking Notes for 20-Minute Presentation

19 Slides | Target Duration: 19-20 minutes

## SLIDE 1: Title Slide

Time: 0:00-0:50 | Duration: 50s

- Introduce yourself: Hubert Kolcz, Warsaw University of Technology
- Hook: ML can fingerprint RNGs even when CHSH tests pass
- Preview: 59% classification accuracy, 20x distinguishability, 3 independent methods validated

## SLIDE 2: CHSH Inequality: Foundation for QKD Security

Time: 0:50-2:10 | Duration: 80s

- CHSH inequality:  $S = \langle AB \rangle + \langle AB' \rangle + \langle A'B \rangle - \langle A'B' \rangle$
- Classical limit:  $S \leq 2$ , Quantum:  $S \leq 2\sqrt{2} \approx 2.828$
- Physical realization: QRNG  $\rightarrow$  FPGA  $\rightarrow$  electro-optical modulator (nanosecond chain)
- Security guarantee:  $S > 2$  certifies device-independent security
- Critical dependency: measurement independence (fresh, unpredictable basis choices)

## SLIDE 3: The Critical Security Gap

Time: 2:10-3:30 | Duration: 80s

Traditional attacks: phase remapping, trojan horse, detector blinding

Our target: RNG entropy source itself

The critical gap:

- DI-QKD proofs account for weak randomness (Santha-Vazirani,  $H_{\infty}=0.186$ )
- BUT: certified once at deployment, then assumed stable
- Min-entropy doesn't capture temporal correlations (Markov biases, autocorrelation)

Our attack exploits monitoring gap:

- RNGs degrade from  $H_{\infty}=0.186$  to  $H_{\infty}=0.05$  without detection
- Unquantified memory violates memoryless assumption
- No continuous validation -> security compromised before detection

Risk: After-the-fact decryption if adversary recorded public transcript

Our contribution: ML framework provides missing continuous monitoring

- Detects min-entropy degradation
- Identifies temporal correlations invisible to static certification
- Markov transitions  $P(1 \rightarrow 1)$ , autocorrelation lag-20, multi-bit patterns

Important: We detect patterns, NOT demonstrated key extraction on real QKD

## SLIDE 4: Multi-Method ML Benchmarking Framework

Time: 3:30-5:00 | Duration: 90s

Three independent methods:

1. qGAN: 12-qubit quantum GAN, KL divergence 0.05-0.20
2. Logistic Regression: linear baseline, 60% accuracy on N=30
3. Neural Network: 30-20-3 architecture, 59% accuracy, replicates from N=3 to N=30

Four-stage operation:

1. Analyze entropy patterns (Markov, autocorrelation, run-length)
2. Correlate hardware metrics (gate fidelity -> Bell correlation  $R^2=0.977$ )
3. Compare platforms (Rigetti 0.8036, IonQ 0.8362, IBM 1.0)
4. Extract fingerprints (device-specific patterns)

Validation design: N=30 with 10 devices per bias class (28 degrees of freedom)

## SLIDE 5: Experimental Methodology & Hardware

Time: 5:00-6:30 | Duration: 90s

Hardware platforms:

- Rigetti Aspen-M-3: 80 qubits, Bell 0.8036, gate fidelity 93.6%
- IonQ Aria-1: 25 qubits, Bell 0.8362, gate fidelity 99.4%
- IBM Qiskit: ideal baseline, Bell 1.0, gate fidelity 100%

Dataset transparency:

- N=3: 6,000 samples from IBMQ noise-injected simulators (DoraHacks 2024)
- NOT actual QPU data—realistic noise models but simulators
- N=30: synthetic devices with controlled bias levels

N=3 fingerprints: Device 0: 54.7% bias,  $P(1 \rightarrow 1)=0.573$  | Device 1: 56.5%,  $P(1 \rightarrow 1)=0.592$  | Device 2: 49.2%,  $P(1 \rightarrow 1)=0.508$

Cross-method validation: Pearson  $r=0.865$  ( $p < 10^{-9}$ ), Spearman  $\rho=0.931$  ( $p < 10^{-14}$ )

## SLIDE 6: Quantitative Analysis: Bit Frequency Distribution

Time: 6:30-7:30 | Duration: 60s

The entropy paradox:

- Device 1: 54.8% bias, entropy 0.986 bits (low bias)
- Device 2: 56.5% bias, entropy 0.979 bits (medium bias)
- Device 3: 59.2% bias, entropy 0.992 bits (HIGH bias, HIGHEST entropy!)

Why paradoxical?

- Device 3 has strongest frequency bias but highest Shannon entropy
- Yet easiest to classify at 70% accuracy
- Shannon entropy = first-order (individual bits)
- ML detects second-order (Markov, autocorrelation, run-length)

Security implication: Passing NIST tests != ML-robust

## SLIDE 7: Markov Chain Transition Matrices

Time: 7:30-8:20 | Duration: 50s

Transition probabilities:

- Device 1:  $P(1 \rightarrow 1) = 0.573$  (moderate persistence)
- Device 2:  $P(1 \rightarrow 1) = 0.592$  (strongest persistence)
- Device 3:  $P(1 \rightarrow 1) = 0.508$  (most symmetric, appears random)

Key finding: Device-specific transition biases create ML fingerprints invisible to NIST tests

Memory attack connection:

- $P(1 \rightarrow 1)$  deviation from 0.5 = statistical memory
- Physical origins: detector afterpulsing, gate errors, thermal drift
- Enables basis prediction with better-than-random probability

## SLIDE 8: Machine Learning Performance Metrics

Time: 8:20-9:50 | Duration: 90s

N=30 Validation (4 panels):

Panel A - Confusion Matrix: 59% overall accuracy, balanced across 3 classes

Panel B - Method Comparison:

- NN: 59%, LR: 60%, Random: 33.3%
- 77% improvement over random baseline

Panel C - N=3 to N=30 Bridging:

- N=3: 58.67% accuracy
- N=30: 59.21% accuracy
- Replicates within 0.54 percentage points
- $p < 10^{-9}$  with 28 degrees of freedom

Panel D - Baseline Improvement: 77-80% above random

Conclusion: Method works on synthetic data with proper statistical power

## SLIDE 9: Neural Network Architecture Analysis

Time: 9:50-10:40 | Duration: 50s

Optimal configuration:

- Batch size: 8 (stable gradients)
- Architecture: 30-20-3 (wider first layer captures more features)
- Regularization: L1  $\lambda=0.002$  (sparse feature selection)
- Training: 1000 epochs necessary

Generalization evidence:

- Optimized on N=3, applied to N=30 without modification
- Performance replicates -> not overfitting specific N=3 noise profiles

## SLIDE 10: Per-Device Classification Performance

Time: 10:40-11:30 | Duration: 50s

Individual results:

- Device 1 (54.8% bias): 66.7% accuracy (moderate)
- Device 2 (56.5% bias): 65.0% accuracy (most challenging—similar to both extremes)
- Device 3 (59.2% bias): 70.0% accuracy (best performance despite highest entropy)

Security takeaway: High entropy necessary but insufficient—need adversarial robustness testing

# SLIDE 11: qGAN Distributional Analysis: Device Distinguishability

Time: 11:30-12:50 | Duration: 80s

Methodology: 435 pairwise KL divergences (all combinations of 30 devices)

Within-class KL (similar devices):

- Class 0-0: mean 0.048, std 0.044
- Class 1-1: mean 0.083, std 0.072
- Class 2-2: mean 0.101, std 0.101

Between-class KL (different classes):

- Classes 0-1: mean 0.670, std 0.369
- Classes 0-2: mean 3.180, std 1.318 (most distinguishable)
- Classes 1-2: mean 0.961, std 0.705

Statistical validation:

- Mann-Whitney U test:  $p=3.26 \times 10^{-6}$
- 20x higher distinguishability between classes
- Cross-method correlation:  $r=0.865$ ,  $\rho=0.931$

## SLIDE 12: Proposed DI-QKD Vulnerability Analysis

Time: 12:50-14:20 | Duration: 90s

### Phase 1: RNG Profiling

- Passive monitoring via side channels
- ML fingerprinting: 59% accuracy (80% above random)
- Bias detection: 59% vs 54% threshold exploitable
- Temporal patterns: Markov  $P(1 \rightarrow 1) = 0.508-0.592$

### Phase 2: Basis Prediction

- Environmental correlation (temperature, gate fidelity)
- CHSH degradation tracking ( $S < 2.2$  signals exploitable noise)
- Basis inference: predict Alice/Bob settings better than random
- Side-channel extraction: entropy deviation + hardware signatures

### Validated foundation:

- Multi-modal: qGAN-NN  $r=0.865$ , between-class KL 20x (both  $p < 10^{-9}$ )
- Hardware:  $R^2=0.977$  (gate fidelity  $\rightarrow$  CHSH  $\rightarrow$  RNG quality)
- Detection threshold: CHSH  $< 2.2$  + bias  $> 59\%$

DI-QKD vulnerability: Security proofs assume stable min-entropy, don't account for temporal correlations

Critical gap: We fingerprint certified QRNGs but haven't demonstrated key extraction on real QKD

## SLIDE 13: Hardware Platform CHSH Correlation Analysis

Time: 14:20-15:20 | Duration: 60s

Platform comparison table:

- IBM: Bell 1.0, gate fidelity 100% (ideal)
- Rigetti: Bell 0.8036, gate fidelity 93.6%
- IonQ: Bell 0.8362, gate fidelity 99.4%

Critical finding:  $R^2=0.977$  correlation (gate fidelity  $\rightarrow$  Bell correlation)

- Gate fidelity predicts certifiable randomness quality
- Easier to measure than full CHSH tests
- Early warning system for RNG vulnerability

Aggregate vs microstructure:

- CHSH verifies aggregate S values (100,000 shots averaged)
- Doesn't verify per-sample microstructure (temporal correlations within 100-bit sequences)
- ML exploits second-order patterns CHSH misses

## SLIDE 14: Statistical Significance & Correlation Analysis

Time: 15:20-16:10 | Duration: 50s

Correlation evidence:

- Pearson:  $r=0.865$  ( $p<10^{-9}$ )
- Spearman:  $\rho=0.931$  ( $p<10^{-14}$ ) — even stronger, non-parametric

Statistical power:  $N=30$  gives  $df=28$ , adequate for medium-to-large effects

NIST test paradox:

- All devices pass chi-square (values  $< 3.841$ )
- All have high Shannon entropy (0.979-0.992 bits)
- Yet NN classifies at 59% with  $p<10^{-9}$

Key message: Passing NIST tests insufficient—need adversarial robustness testing

## SLIDE 15: Proposed Attack Detection Framework

Time: 16:10-17:20 | Duration: 70s

High-quality RNG profile:

- Bell correlation  $\geq 0.8$
- Shannon entropy  $\approx 0.99$  bits
- KL divergence  $\approx 3.7$  (stable baseline)
- Bit frequency: 50%  $\pm 2\%$

Degraded RNG profile (attack signatures):

- Bell correlation  $< 0.8$
- Entropy deviation  $> 5\%$
- KL divergence  $> 17$  (massive shift)
- Bit frequency: 59% (exploitable threshold)

Attack type signatures:

- Phase remapping: correlation drop + entropy oscillation
- Detector blinding: complete loss of quantum correlation
- Temperature attack: gradual bias accumulation
- RNG compromise: persistent frequency bias

Proposed application: Real-time monitoring with multi-indicator thresholds

Critical requirement: Validation on 50+ production QKD RNGs needed

## SLIDE 16: Proposed Application: Metro QKD Security Monitoring

Time: 17:20-18:10 | Duration: 50s

What's validated:

- Framework on  $N=30$  synthetic devices
- 59% accuracy,  $p < 10^{-9}$
- 20x distinguishability,  $p < 10^{-60}$
- Statistical signatures detectable despite passing NIST

Critical gaps for deployment:

1. Need 50+ real production QKD RNGs (not synthetic)
2. Long-term drift monitoring (months/years continuous operation)
3. Demonstration of actual key leakage detection

Honest assessment: We detect patterns, not demonstrated key extraction

## SLIDE 17: Bridging Theory & Engineering Reality

Time: 18:10-19:10 | Duration: 60s

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 certification)
- Environmental factor monitoring (temperature, EMI)
- Hardware drift detection (gradual degradation)
- Real-time attack identification

Future Directions:

- Photonic & topological qubits
- Long-term degradation studies (6-12 months)
- Quantum ML for detection
- NIST/ISO standards development

## SLIDE 18: Comprehensive Validation Summary

Time: 19:10-19:40 | Duration: 30s

Point to comprehensive validation figure (6 panels):

- Top left: bit frequency distributions
- Top right: Markov transitions
- Middle left: qGAN tournament (20x distinguishability)
- Middle right: NN confusion matrix (59% balanced)
- Bottom left: hardware correlation ( $R^2=0.977$ )
- Bottom right: per-device performance (Device 3: 70% despite highest entropy)

Summary message: Multi-method consistency, proper statistical power, N=3 to N=30 replication

## SLIDE 19: Conclusions & Impact

Time: 19:40-21:00 | Duration: 80s

Five key contributions:

1. Device fingerprinting: 59% accuracy on  $N=30$  (73.2% balanced), 120% above baseline,  $r=0.865$ ,  $p<10^{-9}$
2. Multi-method consistency: qGAN-LR-NN converge,  $\rho=0.931$ ,  $p<10^{-14}$
3. qGAN tournament: 20x distinguishability ( $p<10^{-60}$ ), within-class 0.077 vs between-class 1.60
4. Scalability:  $N=3$  58.67% replicates to  $N=30$  59% with strong significance
5. Proposed application: validated on synthetic data, requires real hardware testing

Impact statement: ML-based fingerprinting successfully distinguishes quantum noise profiles on synthetic data

Critical gap reminder: Detecting patterns  $\neq$  exploiting for QKD attacks

- We show RNGs have fingerprints
- Have NOT demonstrated key leakage in production
- Real security impact requires actual key extraction demonstration

References: Zapatero et al. 2023 (Nature npj QI), DoraHacks YQuantum 2024, [github.com/hubertkolcz/NoiseVsRandomness](https://github.com/hubertkolcz/NoiseVsRandomness)

Acknowledgments: Warsaw U Tech, Texas A&M, U Toronto, DoraHacks

Contact: [hubert.kolcz.dokt@pw.edu.pl](mailto:hubert.kolcz.dokt@pw.edu.pl)

Thank you! Questions?