

# Disentangling The “Quantum PUF”: A Multi-Axis Taxonomy of Quantum Lineage, Classical Substrate Hardening, And Attacker Sophistication

Ismoil Makhamatjonov

Department of Computer Science, University of Sheffield, Uzbekistan

Received: 06 Feb 2026 | Received Revised Version: 24 Feb 2026 | Accepted: 12 Mar 2026 | Published: 31 Mar 2026

Volume 08 Issue 03 2026 | Crossref DOI: 10.37547/tajet/Volume08Issue03-14

## Abstract

*Physically unclonable functions (PUFs) convert microscopic manufacturing noise inside a chip into a device-specific hardware fingerprint, but delay-based designs fall to machine-learning modelling once an adversary collects enough challenge-response pairs (CRPs). This article reports an empirical study of six machine-learning attacks — logistic regression, least-squares regression, a multilayer perceptron, an LMN feature-lifting model, a denoising autoencoder, and a compact Transformer encoder — against Arbiter OR-AND-XOR PUFs (AOX-PUFs), simulated with pypuf on the University of Sheffield's Stanage high-performance computing cluster across CRP budgets from 2,500 to 1.9 million. The attacks exploit the Arbiter PUF's linear-additive delay model, and the defence under test is a BB84-inspired encoding wrapper motivated by the Quantum Lock forging-probability bound. Classical AOX-PUFs were modelled with 95–99% accuracy within tens of thousands of CRPs, while BB84-inspired encoding capped every attacker at a flat ceiling of roughly 0.85–0.90 that did not rise within either batch, even at close to two million CRPs. Decomposing the encoding pipeline shows that this ceiling tracks the simulated measurement-success probability rather than a genuine basis-secrecy effect: disclosing the encoding basis altered outcomes by less than a percentage point. An illustrative calculation based on the Quantum Lock bound shows that a genuine quantum advantage should decay the forging probability exponentially as the number of encoded output qubits grows, a prediction the flat empirical ceiling does not fulfil. The results corroborate prior architectural-hardening findings for AOX-PUFs while diverging from the CRP-inflation behaviour the Quantum Lock model predicts.*

**Keywords:** Physical unclonable functions, machine learning attacks, hardware security, quantum-inspired cryptography, BB84 protocol, arbiter PUF, Quantum Lock model.

© 2026 : Ismoil Makhamatjonov. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The authors retain copyright and allow others to share, adapt, or redistribute the work with proper attribution.

**Cite This Article:** Ismoil Makhamatjonov. (2026). Disentangling The “Quantum PUF”: A Multi-Axis Taxonomy of Quantum Lineage, Classical Substrate Hardening, And Attacker Sophistication. The American Journal of Engineering and Technology, 8(03), 181–188. <https://doi.org/10.37547/tajet/Volume08Issue03-14>

## 1. Introduction

A physically unclonable function (PUF) exploits uncontrollable, die-to-die variation in delays and other physical parameters introduced during manufacturing [1], [2]. The challenge-response behaviour functions as a hardware fingerprint, attractive for authentication and key derivation in embedded and IoT systems. The Arbiter PUF (APUF) is the canonical delay-based design: a

challenge steers two signal paths through a chain of switching stages, and an arbiter latch reports which path arrives first. Because the response is close to the sign of a linear combination of the delay parameters, APUFs are highly vulnerable to modelling: an adversary who harvests a modest number of clean CRPs predicts the response function with high accuracy using ordinary supervised learning [3], [4]. Deep neural networks now

model even deliberately hardened compositions such as XOR-APUFs at scale [5]; response-bias analysis reveals exploitable structure beyond raw accuracy arguments [6]; and evolutionary strategies exploit reliability information to break designs that resist gradient-based learners [7].

The hardware-security community's principal response has been architectural: combine several Arbiter PUFs with non-linear logic to complicate the learnable decision surface. The Arbiter OR-AND-XOR PUF (AOX-PUF, also OAX-PUF) pools sub-PUF responses through parallel OR, AND, and XOR blocks of relative sizes ( $x$ ,  $y$ ,  $z$ ), then combines them with a further XOR [8]. A formal uniformity analysis shows the composition is unbiased — response uniformity of exactly 0.5 whenever the XOR block is non-empty ( $z \geq 1$ ) — even though its OR and AND sub-blocks are individually biased [8]. This hardening slows attackers down but does not stop them: given sufficient clean CRPs, high-capacity learners continue to progress [8], [5]. A complementary line of research argues that PUF security is governed less by circuit complexity than by how many CRPs an adversary observes and how clean they are. BB84-style encoding targets this factor by embedding a classical PUF's output into non-orthogonal quantum states; an eavesdropper who does not know the encoding basis is forced by the no-cloning theorem to measure in a possibly wrong basis, reducing the exploitable information per CRP [9]. The resulting Quantum Lock model predicts that a basis-unknown attacker needs substantially more CRPs than a basis-known one to reach a target accuracy, and that this advantage collapses once the basis leaks [10].

Delay-based PUFs continue to be deployed for low-cost authentication despite this well-documented vulnerability [3]–[7], and architectural hardening alone has repeatedly been shown to raise rather than remove the ceiling on classical attacker accuracy [5], [8]. The Quantum Lock bound [10] has rarely been tested against a modern battery of attackers that includes deep architectures such as autoencoders and Transformer encoders alongside the classical baselines the theoretical literature assumes. This article addresses that gap for one hybrid architecture — a BB84-wrapped AOX-PUF — across six attacker families and three orders of magnitude of CRP budget, testing two competing hypotheses. H1 (Quantum Lock): BB84-inspired encoding reproduces the predicted CRP-inflation effect, so basis-unknown

attackers need substantially more CRPs than basis-known ones to reach a fixed accuracy target, with the gap widening as the budget grows. H0 (measurement-noise): any ceiling under encoding is an artefact of the simulator's measurement-success probability,  $p_{\text{succ}}$ , and is largely insensitive to whether the basis is later disclosed. Sections 3 and 4 test H1 against H0, using the short-budget (2,500–75,000 CRP) and long-budget (up to 1.9 million CRP) batches of Section 2.

## 2. Methods

### 2.1 Simulation infrastructure and experimental plan

All experiments used the pypuf simulation framework [11], extended with a custom BB84 encoding wrapper, and ran on the University of Sheffield's Stanage HPC cluster, in batches of 32 nodes at 128 GB of memory each — approximately 4 TB of aggregate memory for the largest runs. AOX-PUFs were instantiated with total fan-in of 5, 6, and 7 (families aox-5, aox-6, aox-7), each built from independent Arbiter PUF sub-blocks combined through OR, AND, and XOR branches, with a noise parameter modelling operational instability. Twelve ( $x$ ,  $y$ ,  $z$ ) compositions were evaluated across the three families, including configurations with no XOR branch, to probe the boundary between reliable and unlearnable behaviour. Six attackers were evaluated against every configuration: logistic regression and least-squares regression as linear baselines consistent with the delay model of Eq. (1); a multilayer perceptron and an LMN feature-lifting model as established pypuf attacks; and two attackers built for this project — a denoising autoencoder with a classification head, and a compact Transformer encoder treating the challenge as a bit sequence. Two batches ran at different scales: a short-budget sweep from 2,500 to 75,000 CRPs across all six attackers ( $p_{\text{succ}} \approx 0.85$ ), and a long-budget sweep from about 1 million to 1.9 million CRPs focused on logistic regression — the strongest, cheapest classical attacker — with  $p_{\text{succ}} \approx 0.90$ .

### 2.2 AOX-PUF architecture and the linear-additive delay model

Each sub-block is an Arbiter PUF, whose behaviour follows the standard linear-additive delay model [3], [12]. For a challenge  $c = (c_0, \dots, c_{n-1})$ , the delay difference between the two arbiter paths is

$$\Delta = w_0 \Psi_0 + w_1 \Psi_1 + \dots + w_n \Psi_n = \sum_{i=0}^n w_i \Psi_i = \langle w, \Psi \rangle \quad (1)$$

where  $w$  is a vector of unknown weights fixed by the manufacturing variation of the specific chip, and  $\Psi$  is a parity (feature) vector derived deterministically from the challenge, with  $\Psi_n = 1$  and  $\Psi_i = \prod_{j=i}^{n-1} (1 - 2c_j)$ , for  $i = 0, \dots, n - 1$ . The response bit is obtained by thresholding  $\Delta$  at zero ( $r = \text{sgn}(\Delta) \in \{-1, +1\}$ , mapped to a bit as  $\frac{1-r}{2}$ ). Equation (1) is the formal reason this PUF family is called “mathematically clonable”: the response is a linear function of a feature vector  $\Psi$  an attacker can compute directly from the challenge. Recovering  $w$  from observed  $(\Psi, r)$  pairs is exactly the problem logistic regression — one of the six attackers benchmarked here — is designed to solve, and this correspondence explains why every Arbiter-PUF-based design, including the AOX-PUF evaluated here, is modelled with high accuracy once an adversary has sufficient clean CRPs (Section 3.1).

### 2.3 Uniformity of the OAX composition

The AOX-PUF combines sub-PUF outputs through OR, AND, and XOR blocks of sizes  $x$ ,  $y$ , and  $z$ , and a further XOR across the three blocks. Yao et al. address formally whether this non-linear combination biases the response distribution [8]. Writing response uniformity as the proportion of 1-valued response bits over  $n$  measurements,

$$\text{Uniformity} = \frac{1}{n} \sum_{i=1}^n r_i \times 100\% \quad (2)$$

they show that an  $x$ -OR-PUF is biased toward 1 with  $U_{\text{or}} = 1 - \frac{1}{2^x}$  and a  $y$ -AND-PUF is biased toward 0 with  $U_{\text{and}} = \frac{1}{2^y}$ . The full  $(x, y, z)$ -OAX-PUF nevertheless has

$$p_{\text{forge}}^{\text{quantum}} \leq p_{\text{extract}} \times p_{\text{forge}}^{\text{classical}} \quad (3a)$$

where

$$p_{\text{extract}} \leq \sum_{k=\lceil(1-\varepsilon)q\rceil}^q \binom{q}{k} p_{\text{guess}}^{2mk} (1 - p_{\text{guess}}^{2m})^{q-k} \quad (3b)$$

Here  $q$  is the number of CRPs available to the adversary,  $m$  is the number of output qubits the HPUF produces per query (each qubit encodes two bits of the classical PUF output, which gives the exponent  $2m$ ),  $\varepsilon \in [0, 1]$  is a tolerance parameter governing how many of the  $q$  extractions must succeed, and  $p_{\text{guess}} \in [0.5, 1]$  is the probability of correctly guessing a single bit of the classical output from a single non-orthogonal qubit measurement. The theorem's central claim is that, for sufficiently small  $\varepsilon$ ,  $p_{\text{extract}}$  decays exponentially as  $m$  increases, so a defender can, in principle, drive the quantum forging probability arbitrarily below the classical one simply by encoding more output bits per

uniformity exactly  $1/2$  whenever  $z \geq 1$ , provided each sub-PUF has uniformity  $0.5$ , because XOR-ing any bit with an independent unbiased bit yields an unbiased bit. When  $z = 0$ , the response is the XOR of the OR and AND blocks, with uniformity  $U_{\text{or}}(1 - U_{\text{and}}) + (1 - U_{\text{or}})U_{\text{and}}$ , which is generally not  $1/2$  (it is  $0.6875$  for  $x = 3$ ,  $y = 2$ ). This matters methodologically: the AOX-PUF baseline attacked here is, by construction, an unbiased binary target for every configuration with  $z \geq 1$  before any BB84 encoding is applied, so attacker accuracy above  $0.5$  at low CRP budgets reflects genuine learning rather than an exploitable label imbalance.

### 2.4 BB84 encoding wrapper and the Quantum Lock bound

Three CRP views were generated per configuration: classical (clean CRPs); BB84 state-only, where the attacker never learns the encoding basis; and BB84 basis-known, where the basis is disclosed partway through collection to simulate a protocol leak. A challenge is XORed with a random basis mask before the PUF evaluates it, and the resulting response bit is flipped independently with probability  $(1 - p_{\text{succ}})$ , modelling a realistic measurement-success rate.

This wrapper is directly inspired by, and intended as an empirical test of, the security argument Chakraborty et al. give for quantum-encoded PUFs [10]. Their theorem upper-bounds the probability of forging a quantum-encoded PUF (a Hybrid Locked PUF, HPUF) in terms of an extraction probability and the forging probability of the underlying classical PUF:

query — the behaviour H1 expects to see in the empirical ceiling, which Section 4.2 compares with the ceiling actually measured.

## 3. Results

### 3.1 Short-budget batch

At the short-budget scale, classical AOX-PUFs modelled with high accuracy well within the tested range. Logistic regression climbed to approximately  $0.95$ – $0.96$  by  $25,000$ – $50,000$  CRPs across most  $(x, y, z)$  mixes, and the multilayer perceptron, least-squares and autoencoder attackers reached comparable or higher accuracy (up to  $\approx 0.98$ – $0.99$ ). The Transformer and LMN attackers stayed

close to the 0.5 uniformity baseline of [Section 2.3](#) ( $\approx 0.51$ – $0.60$ ) for much of this range, consistent with slower convergence rather than label imbalance. Every configuration's BB84-encoded traces — state-only and

basis-known alike — remained essentially flat around  $0.84$ – $0.85$  across the entire range, regardless of attacker ([Table 1](#)).

| Attack model             | Classical accuracy   | AOX-PUF | BB84 state-only (flat ceiling) | BB84 basis-known (flat ceiling) |
|--------------------------|----------------------|---------|--------------------------------|---------------------------------|
| Logistic Regression      | 0.55–0.99            |         | $\sim 0.85$ – $0.90$           | $\sim 0.85$ – $0.90$            |
| LMN (feature-lifting)    | $\sim 0.52$ – $0.60$ |         | $\sim 0.85$                    | $\sim 0.85$                     |
| Multilayer Perceptron    | 0.56–0.98            |         | $\sim 0.85$                    | $\sim 0.85$                     |
| Least-Squares Regression | 0.51–0.98            |         | $\sim 0.85$                    | $\sim 0.85$                     |
| Autoencoder              | $\sim 0.99$          |         | $\sim 0.85$                    | $\sim 0.85$                     |
| Transformer              | $\sim 0.51$ – $0.56$ |         | $\sim 0.85$                    | $\sim 0.85$                     |

**Table 1. Accuracy bands by attack model: classical AOX-PUF vs. BB84-encoded ceiling, short-budget batch.**

### 3.2 Long-budget batch

Extending logistic regression to the long-budget regime (up to 1.9 million CRPs) confirmed the short-budget pattern. Classical accuracy plateaued between approximately 0.92 and 0.99 depending on circuit mix, while both BB84-encoded traces stayed pinned at approximately 0.90 across the full 1–1.9 million CRP range; nearly doubling the data gave the attacker no measurable improvement. [Table 2](#) reports this gap per configuration, ranging from  $\approx 0.02$  (all-XOR (0,0,5)) to

$\approx 0.08$ – $0.09$  ((2,2,1) and aox-6 (3,2,1), the largest): topology shapes how far the classical ceiling rises above the encoded floor, but does not move the floor itself. One aox-5 mix with  $(x, y, z) = (3, 2, 0)$  and no XOR branch returned an accuracy of  $\approx 0.00$  ( $\approx 0.10$  under encoding). A binary classifier at 0% accuracy is a perfect predictor with inverted labels, whereas a constant-output model would score near the majority-class rate ( $\approx 0.69$  for this mix; [Section 2.3](#)). The result therefore most likely reflects a label-inversion error in the evaluation pipeline and is excluded from further analysis pending a re-run.

| Configuration  | Classical (LR)              | Q-state (LR) | Q-basis (LR) | Classical–encoded gap |
|----------------|-----------------------------|--------------|--------------|-----------------------|
| aox-5, (0,0,5) | $\sim 0.92$ – $0.93$        | $\sim 0.90$  | $\sim 0.90$  | $\sim 0.02$           |
| aox-5, (2,1,2) | $\sim 0.96$ – $0.97$        | $\sim 0.90$  | $\sim 0.90$  | $\sim 0.06$           |
| aox-5, (2,2,1) | $\sim 0.98$ – $0.99$        | $\sim 0.90$  | $\sim 0.90$  | $\sim 0.08$ – $0.09$  |
| aox-5, (3,2,0) | $\sim 0.00$<br>(degenerate) | $\sim 0.10$  | $\sim 0.10$  | artefact, $z = 0$     |
| aox-6, (2,2,2) | $\sim 0.96$ – $0.97$        | $\sim 0.90$  | $\sim 0.90$  | $\sim 0.06$ – $0.07$  |
| aox-6, (3,2,1) | $> 0.98$                    | $\sim 0.90$  | $\sim 0.90$  | $\sim 0.08$ – $0.09$  |

**Table 2. Accuracy by  $(x, y, z)$  configuration, logistic regression, long-budget batch ( $\approx 1$ – $1.9$  million CRPs).**

### 3.3 Basis disclosure and attacker-architecture comparison

Across both batches, the state-only and basis-known BB84 traces stayed within a fraction of a percentage point of one another at every tested budget, for every attacker and configuration, with no pair showing a widening separation as the budget grew. Consistent with the correspondence between logistic regression and the linear delay model of [Eq. \(1\)](#), logistic regression reached its ceiling with fewer CRPs than most non-linear attackers, although every attacker eventually converged to the same encoded value ( $\approx 0.85$  in the short-budget batch). The denoising autoencoder was the fastest classical learner overall, reaching its own  $\approx 0.99$  ceiling at smaller budgets than logistic regression, yet it was capped at the identical  $\approx 0.85$  ceiling once BB84 encoding was applied — an early sign, confirmed in [Section 4](#), that the encoded ceiling is a property of the protocol rather than of any attacker's architecture or capacity.

## 4. Discussion

### 4.1 Why the ceiling is flat, and the effect of basis disclosure

A more important question is why the encoded curves are flat. The simulator encodes a challenge with a random basis before the PUF evaluates it, then flips the resulting bit independently with probability  $(1 - p_{succ})$  to model measurement noise. Because this label noise is independent of the attacking model and fixed by  $p_{succ}$ , no learner, however capable, can exceed  $p_{succ}$  when evaluated on held-out labels that carry the same noise. More precisely, a model with clean-label accuracy  $A_{cls}$  is expected to score  $A_{enc} = p_{succ} \cdot A_{cls} + (1 - p_{succ})(1 -$

$A_{cls})$  on flipped labels, which equals  $p_{succ}$  only when  $A_{cls} = 1$ . The observed ceiling is therefore a statistical noise ceiling, not a demonstration of computational hardness. Consistent with this, the ceiling rose from  $\approx 0.85$  in the short-budget batch to  $\approx 0.90$  in the long-budget batch, matching the change in  $p_{succ}$  from 0.85 to 0.90.

This has a direct, testable consequence already reported in [Section 3.3](#): revealing the encoding basis had essentially no effect. Under H1, the two traces should separate increasingly as the budget grows, since basis secrecy is the mechanism the theorem credits with suppressing  $p_{extract}$ ; under H0, they should stay close throughout, since basis knowledge changes which features the attacker observes but not the fixed label-flip probability that bounds accuracy. Near-identical traces from the first measured budget onward, with no widening gap even at 1.9 million CRPs, favour H0. In the theoretical model the basis is the security-critical secret [\[10\]](#); here it was not the dominant factor.

### 4.2 Illustrative calculation: the Quantum Lock bound predicts exponential decay with m

The divergence between H1 and H0 can be made concrete with the Quantum Lock bound of [Eq. \(3\)](#) in a small, illustrative example. The calculation below is illustrative only: it uses the published theoretical bound to show what theory predicts as a defender adds more encoded output qubits ( $m$ ), and is not an experimental result of this study, since the pipeline implemented here never varied  $m$ .

Take the simplest non-trivial case of [Eq. \(3b\)](#): a single CRP ( $q = 1$ ) with an exact-match tolerance ( $\varepsilon = 0$ , so  $k$  ranges only over  $k = q = 1$ ). The summation then collapses to a single term:

$$p_{extract} \leq \binom{1}{1} p_{guess}^{2m} (1 - p_{guess}^{2m})^0 = p_{guess}^{2m} \quad (4)$$

For the guessing probability, take  $p_{guess} = 0.75$  — a value in the plausible range for a single-qubit, non-orthogonal-state guess, chosen purely for illustration and not calibrated to this study's measured ceiling of  $\approx 0.85$ – $0.90$ ; [Section 4.1](#) already establishes that the measured ceiling arises from a different mechanism ( $p_{succ}$ ), not from  $p_{guess}$ .

For  $m = 2$  encoded qubits:  $p_{extract} \leq (0.75)^{2 \times 2} = (0.75)^4 = 0.5625^2 = 0.31640625$ , i.e. the bound

allows roughly 31.6% residual extraction probability. For  $m = 8$  encoded qubits:  $p_{extract} \leq (0.75)^{2 \times 8} = (0.75)^{16}$ . Computing stepwise:  $(0.75)^2 = 0.5625$ ;  $(0.75)^4 = 0.31640625$ ;  $(0.75)^8 \approx 0.10011$ ;  $(0.75)^{16} \approx 0.01002$ , i.e. the bound allows only about 1.0% residual extraction probability.

Quadrupling the number of encoded qubits from  $m = 2$  to  $m = 8$  therefore drives the upper bound on  $p_{extract}$  down by a factor of roughly 31 (from  $\approx 31.6\%$  to  $\approx 1.0\%$ ),

and, via [Eq. \(3a\)](#), would drive the bound on  $p_{\text{forge}}^{\text{quantum}}$  down by the same factor relative to  $p_{\text{forge}}^{\text{classical}}$  — the exponential-decay-with- $m$  behaviour that gives the Quantum Lock construction its theoretical advantage over an unprotected classical PUF.

No equivalent decay appears in the empirical ceiling reported in [Section 3](#): across CRP budgets from 2,500 to 1.9 million, and regardless of whether the encoding basis was disclosed, the measured ceiling stayed within a narrow band of  $\approx 0.85$ – $0.90$  throughout, because the wrapper evaluated here never varied  $m$ , nor implemented the adaptive extraction procedure [Eq. \(3b\)](#) formalises — it applied a single, fixed label-flip probability after PUF evaluation. For H1 to hold, the ceiling would have to fall geometrically as encoded qubits are added. The flat, basis-insensitive ceiling measured here therefore provides no evidence of the Quantum Lock mechanism at work; it reflects a simpler effect that happens to produce a similar flat accuracy cap.

### 4.3 Comparison with the global literature

[Table 3](#) summarises where the results agree with prior work and where they depart from it. They confirm Yao et al.'s finding that AOX/OAX recomposition raises the bar for linear attackers without closing it [\[8\]](#): logistic regression reached comparable accuracy bands at similar CRP scales, and the degenerate zero-XOR configuration echoes their caution that a single-gate branch can become redundant, explained formally by the uniformity result of [Eq. \(2\)](#). This study extends that architecture-only evaluation with a protocol-level layer on the same substrate: topology governs the classical ceiling ([Table 2](#)), while  $p_{\text{succ}}$  governs the encoded ceiling, with little

interaction between the two. The Interpose PUF of Nguyen et al. [\[12\]](#) is an alternative architectural strategy for the same problem; both designs are, in the sense of [Eq. \(1\)](#), linear-delay compositions combined through different non-linear functions, and logistic regression remained among the strongest attackers here exactly as that shared substrate predicts. Wisiol and Pirnay's demonstration that XOR-APUFs carry exploitable response bias beyond raw CRP-scaling [\[6\]](#) cautions that [Eq. \(2\)](#) rules out aggregate bias only, not the finer-grained conditional bias they identify, which this study did not test directly.

Where this study diverges most clearly is from Chakraborty et al.'s own prediction of a widening, basis-sensitive separation [\[10\]](#) and the exponential-decay-with- $m$  behaviour [Eq. \(3b\)](#) predicts, concretised in [Section 4.2](#): the two basis-regime traces converge to an identical flat ceiling almost immediately and never move, because the pipeline never varied  $m$ . This is best read as a property of how BB84 encoding was operationalised — a fixed label-flip probability rather than the theorem's adaptive extraction procedure — not a refutation of the Quantum Lock result, following the reporting practice Arapinis et al. recommend for quantum-PUF proposals [\[9\]](#). Fei et al. [\[13\]](#), Zhuang et al. [\[14\]](#), and Millwood et al. [\[16\]](#) pursue the same goal — throttling the challenge-response interface rather than the circuit — through mechanisms that, as [Table 3](#) shows, produce different accuracy-versus-CRP curves, consistent with the modelling-attack literature following Rührmair et al. [\[3\]](#), [\[4\]](#) and recent evidence that deep learning continues to erode PUF hardening at scale [\[5\]](#).

| Prior work             | Prior claim / finding                                                                                                        | This study's finding                                                                                            |
|------------------------|------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Yao et al. (2022)      | AOX/OAX raises the bar for linear attackers but does not close it; $U = 0.5$ regardless of $(x, y, z)$ <a href="#">[8]</a> . | LR reached 0.92–0.99 at similar CRP scales; the (3,2,0) mix was degenerate — confirms directly.                 |
| Nguyen et al. (2019)   | Interpose PUF formalises the linear delay model as an alternative hardening strategy <a href="#">[12]</a> .                  | LR remained among the strongest classical attackers on every mix tested — confirms the shared linear substrate. |
| Wisioł & Pirnay (2020) | XOR-APUFs carry exploitable response bias beyond raw CRP-count scaling <a href="#">[6]</a> .                                 | Not tested directly; <a href="#">Eq. (2)</a> rules out aggregate bias only — untested confound, future work.    |

|                                            |                                                                                                              |                                                                                                                                                                    |
|--------------------------------------------|--------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Chakraborty et al. (2023)                  | $p_{\text{extract}}$ decays exponentially with encoded qubits $m$ ; a basis-sensitive gap should widen [10]. | State-only and basis-known traces stay nearly identical; the pipeline never varied $m$ — diverges (Section 4.2).                                                   |
| Fei (2024); Zhuang (2021); Millwood (2023) | Optimal hybrid-PUF attacks [13]; classical challenge obfuscation [14]; DRAM-based obfuscation [16].          | Unencoded AOX-PUF fell fastest to logistic regression/autoencoder; BB84 imposes a hard ceiling rather than delaying it — complementary, different defence profile. |

**Table 3. Comparison between prior findings in the literature and this study's results.**

#### 4.4 Implications and limitations

The relevant question for evaluators is what accuracy ceiling a protocol creates, not merely whether attacks become harder: capping accuracy at 90% is a materially different claim from raising the CRP budget needed to reach 99% by an order of magnitude. Here, basis secrecy contributed almost nothing beyond the measurement noise already present in both regimes, visible only once both regimes are measured side by side. Circuit topology strongly shaped classical difficulty but had almost no effect once BB84 encoding was active, so architectural hardening and protocol-level throttling should be evaluated independently rather than assumed to compound. This threat model assumed direct access to clean CRPs; Fei et al.'s analysis of attacks under minimal adversarial knowledge [15] shows that partial or noisy observations can still yield useful model information.

This study is simulation-based, so a fabricated silicon prototype is needed to confirm these dynamics under genuine measurement noise. Future work should treat  $p_{\text{succ}}$  as an experimental variable, sweep  $m$  directly rather than only illustratively, test the encoding against reliability-based [7] and minimal-knowledge [15] attackers, and evaluate whether combining classical [14] or DRAM-based [16] obfuscation with BB84-style encoding closes the gap to the exponentially-decaying ceiling the Quantum Lock bound predicts.

#### 5. Conclusion

Six machine-learning attackers, evaluated across three orders of magnitude of CRP budget against AOX-PUFs on a university HPC cluster, confirm that unprotected delay-based PUFs succumb to modelling within a modest data budget, reaching 95–99% accuracy within tens of thousands of CRPs — as the linear-additive delay model of Eq. (1) predicts for a design whose response is a linear function of a computable feature vector. Wrapping the

challenge-response interface in BB84-inspired encoding capped every attacker at a flat ceiling of approximately 0.85–0.90 that did not shift even near two million CRPs, regardless of circuit topology. The central finding concerns the mechanism behind that ceiling rather than its existence: it tracked simulated measurement noise rather than the basis-secrecy property the Quantum Lock bound credits with suppressing extraction probability, since disclosing the basis altered outcomes negligibly and the illustrative calculation of Section 4.2 shows the predicted exponential-decay-with- $m$  behaviour is absent from the flat ceiling actually measured. The evidence supports  $H_0$  rather than  $H_1$ , and this distinction should inform how such hybrid PUF defences are specified and evaluated.

#### References

1. B. Gassend, D. Clarke, M. van Dijk, and S. Devadas, "Silicon physical random functions," in Proc. 9th ACM Conf. Comput. Commun. Secur., 2002, pp. 148–160.
2. C. Herder, M.-D. Yu, F. Koushanfar, and S. Devadas, "Physical unclonable functions and applications: A tutorial," Proc. IEEE, vol. 102, no. 8, pp. 1126–1141, 2014.
3. U. Rührmair, F. Sehnke, J. Sölter, G. Dror, S. Devadas, and J. Schmidhuber, "Modeling attacks on physical unclonable functions," in Proc. 17th ACM Conf. Comput. Commun. Secur., 2010, pp. 237–249.
4. U. Rührmair et al., "PUF modeling attacks on simulated and silicon data," IEEE Trans. Inf. Forensics Security, vol. 8, no. 11, pp. 1876–1891, 2013.
5. H. Wang, W. Hao, Y. Tang, B. Zhu, W. Dong, and W. Liu, "Deep neural network modeling attacks on arbiter-PUF-based designs," Cybersecurity, vol. 8, no. 11, 2025, doi: 10.1186/s42400-024-00308-7.

6. N. Wisiol and N. Pirnay, "XOR arbiter PUFs have systematic response bias," in *Financial Cryptography and Data Security (FC 2020)*, LNCS vol. 12059, Springer, 2020, pp. 50–57, doi: 10.1007/978-3-030-51280-4\_4.
7. G. T. Becker, "The gap between promise and reality: On the insecurity of XOR arbiter PUFs," in *Cryptographic Hardware and Embedded Systems – CHES 2015*, LNCS vol. 9293, Springer, 2015, pp. 535–555.
8. J. Yao, L. Pang, Y. Su, Z. Zhang, W. Yang, A. Fu, and Y. Gao, "Design and evaluate recomposited OR-AND-XOR-PUF," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 2, pp. 662–677, 2022, doi: 10.1109/TETC.2022.3170320.
9. M. Arapinis, M. Delavar, M. Doosti, and E. Kashefi, "Quantum physical unclonable functions: Possibilities and impossibilities," *Quantum*, vol. 5, art. 475, 2021.
10. K. Chakraborty, M. Doosti, Y. Ma, C. Wadhwa, M. Arapinis, and E. Kashefi, "Quantum Lock: A provable quantum communication advantage," *Quantum*, vol. 7, art. 1014, 2023.
11. N. Wisiol, C. Gräbnitz, C. Mühl, B. Zengin, T. Soroceanu, N. Pirnay, K. T. Mursi, and A. Baliuka, "pypuf: Cryptanalysis of physically unclonable functions," 2021. [Software]. Available: <https://github.com/nils-wisiol/pypuf>
12. P. H. Nguyen, D. P. Sahoo, C. Jin, K. Mahmood, U. Rührmair, and M. van Dijk, "The interpose PUF: Secure PUF design against state-of-the-art machine learning attacks," *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, vol. 2019, no. 4, pp. 243–290, 2019, doi: 10.13154/tches.v2019.i4.243-290.
13. H. Fei, P. Gope, O. Millwood, and B. Sikdar, "Optimal machine-learning attacks on hybrid PUFs," in *Computer Security – ESORICS 2024*, LNCS vol. 14982, Springer, 2024, pp. 251–270, doi: 10.1007/978-3-031-70879-4\_13.
14. Y. Zhuang, K. T. Mursi, and L. Gaoxiang, "A challenge obfuscating interface for arbiter PUF variants against machine learning attacks," *arXiv:2103.12935*, 2021.
15. H. Fei, O. Millwood, P. Gope, J. Miskelly, and B. Sikdar, "Attacking delay-based PUFs with minimal adversarial knowledge," *IEEE Trans. Inf. Forensics Security*, 2024, arXiv:2403.00464.
16. O. Millwood, M. Kurt Pehlivanoglu, A. Mohammadi Pasikhani, J. Miskelly, P. Gope, and E. B. Kavun, "A generic obfuscation framework for preventing ML-attacks on strong-PUFs through exploitation of DRAM-PUFs," in *2023 IEEE 8th European Symp. Security and Privacy (EuroS&P)*, 2023, doi: 10.1109/EuroSP57164.2023.00015.