Quantum Dropout Does Not Reliably Defend: Adaptive Tests in Quantum Neural Networks

Pablo Rivas

Department of Computing Technology, School of Computer Science and Mathematics, Marist University, 3399 North Road, Poughkeepsie, NY 12601, USA
Pablo.Rivas@Marist.edu

Abstract. Quantum computers may support learning systems whose decisions are read from measured quantum states. Like classical models, quantum classifiers can be vulnerable to small, deliberate input changes. Random prediction rules may obstruct an attack, but can also hide weakness unless the attacker accounts for their randomness. Here we test quantum rotation dropout, which randomly suppresses trainable circuit rotations, as a candidate protection for small quantum neural networks. Here we show that, when attacks account for this random prediction rule, rotation dropout does not provide a consistent robustness benefit in the circuits and binary tasks studied. We compare conventionally trained and adversarially trained classifiers, with and without dropout, under adaptive and transferred attacks, while recording the circuit-evaluation cost of valid testing. Dropout does not improve adversarial training in any required comparison, whereas adversarial training without dropout benefits selected variational classifiers. A separate limited confirmation check does not change the main result. These findings set a practical standard for secure quantum learning: randomized protections should be tested against the prediction rule that is actually deployed, and future defenses should be trained for that test.

Keywords: quantum machine learning · quantum neural networks · adversarial robustness · quantum dropout · adaptive attacks · trustworthy artificial intelligence

1 Introduction

Quantum neural networks (QNNs) provide compact trainable circuits for classification, but a useful classifier must be assessed under perturbed as well as clean inputs. Prior work shows that quantum learning models can be vulnerable to adversarial input perturbations and that QML robustness requires attack-based evaluation [5,10]. This is critical when a proposed protection changes the deployed prediction rule: a stochastic method may appear less vulnerable to an attack that ignores its randomness.

Quantum dropout is a clear test case. It was introduced as QNN regularization intended to reduce overfitting by randomly removing trainable operations during training [8]. Applied to learned rotation gates, dropout could plausibly reduce

reliance on individual angles. Yet plausibility is not robustness: randomized classifiers require attacks that estimate gradients through randomness, such as expectation over transformation (EOT) [1]. QML noise-layer and randomized-encoding defenses have already been studied in adversarial settings [4,3]. Targeted literature screening did not identify a close prior report of trainable rotation dropout under the combined ordinary, adaptive, transferred, and adversarial-training comparison used here. We therefore ask: does *trainable rotation dropout* remain useful when its stochastic inference rule is included in the attack?

We answer this question using exact-state simulation of small QNN classifiers on two generated binary tasks. The required matrix compares a one-wire data-reuploading classifier and a two-wire variational classifier under standard, rotation-dropout, adversarial, and adversarial-plus-dropout training. Deterministic models are assessed with projected-gradient attacks; stochastic inference is assessed with EOT-PGD and Monte Carlo prediction. We also report transferred attacks, execution-count cost, and a four-wire Circle check. The purpose is to determine whether rotation dropout earns a defense claim under attack-aware testing, not to assume that it does, leading to the following research questions:

RQ1. Does adaptive testing support rotation dropout over standard training?
RQ2. Does dropout improve PGD adversarial training?
RQ3. Do transfer tests and the four-wire check agree with the adaptive result?
RQ4. What QNode cost does adaptive stochastic testing require?

This paper has the following contributions:

- a fixed PGD, EOT-PGD, transfer, and adversarial-training evaluation of trainable rotation dropout in small simulated QNNs;
- negative security evidence: no consistent adaptive dropout benefit, and no required-cell mean gain when dropout is added to adversarial training; and
- supporting transfer and four-wire checks plus QNode execution-count accounting for fair stochastic evaluation.

All claims are limited to the executed simulator protocol, selected models and rates, and five paired seeds; they exclude hardware, finite-sampling, scaling, and general impossibility statements.

2 Quantum and Adversarial Background

A quantum neural network (QNN) is used here as a probabilistic binary classifier: a classical point is encoded into a small parameterized circuit, the circuit is measured, and its output probabilities determine a label. All evidence in this paper comes from exact-state simulation; no claim about execution on a quantum processor is made.

2.1 From a qubit to a classifier

A qubit is a normalized two-level state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where $|\alpha|^2 + |\beta|^2 = 1$. Rotation gates change these amplitudes; some angles encode an input and others are learned during training. For a binary classifier with circuit $U(x, \theta)$, measurement defines the class probabilities

$$|\psi(x;\theta)\rangle = U(x,\theta)|0\rangle^{\otimes n}, \qquad p_\theta(c \mid x) = \langle\psi(x;\theta)|\Pi_c|\psi(x;\theta)\rangle, \tag{1}$$

where x is a classical feature vector, θ contains learned angles, n is the number of wires, and Π_c is the measurement operator for class c.

A small circuit receives limited information from one encoding of x. Data re-uploading alternates input-encoding blocks with learned processing blocks,

$$U(x,\theta) = W_L(\theta_L)S(x)\cdots W_1(\theta_1)S(x), \tag{2}$$

where $S(x)$ encodes the features and $W_\ell(\theta_\ell)$ is trainable. Pérez-Salinas et al. introduced this repeated-encoding design for quantum classification [7]. The Methods section specifies the one-wire re-uploading model and two-wire variational model used in our test.

2.2 Rotation dropout and random predictions

Scala et al. introduced quantum dropout as a QNN regularization method [8]. Our tested variant masks only learned rotation angles. For angle θ_j and dropout probability p, we have that:

$$m_j \sim \text{Bernoulli}(1-p), \qquad \widetilde{\theta}_j = m_j\theta_j, \qquad \overline{p}_\theta(c \mid x) = \frac{1}{M}\sum_{r=1}^{M} p_{\theta,m^{(r)}}(c \mid x). \tag{3}$$

When $m_j = 0$, that learned rotation has angle zero for the sampled circuit; input encodings and fixed coupling operations remain active. The last term is the Monte Carlo prediction used when masks are sampled at inference. Random prediction can change accuracy or calibration, but it is not by itself a security guarantee.

2.3 Adversarial inputs and informed testing

Prior QML work has shown that quantum classifiers can be vulnerable to crafted input perturbations [5]. In an untargeted bounded attack, the attacker searches for an input $x_{\text{adv}} = x + \delta$ that increases classification loss while satisfying

$$\|\delta\|_\infty \le \epsilon, \qquad x_{\text{adv}} \in [-1,1]^2. \tag{4}$$

Clean accuracy evaluates unmodified points, while robust accuracy evaluates points after attack. Attack success rate counts originally correct points that become incorrect after attack.

For stochastic inference, an attack against one mask may understate vulnerability. Athalye et al. established that a randomized classifier should be attacked by

estimating its gradient through the random process [1]. In this paper, expectation over transformation (EOT) estimates that gradient as

$$\nabla_x \mathbb{E}_m[\mathcal{L}(p_{\theta,m}(\cdot \mid x), y)] \approx \frac{1}{K} \sum_{k=1}^{K} \nabla_x \mathcal{L}(p_{\theta,m^{(k)}}(\cdot \mid x), y). \tag{5}$$

Hence the paper asks a strict question: does rotation dropout remain useful when its randomness is included in the attack? A negative result is valuable because it prevents an unverified randomized rule from being credited as a defense and sets a clear standard for later QML robustness studies.

3 Related Work

Small quantum classifiers and dropout. The one-wire model in this study is based on data re-uploading, where classical features are inserted repeatedly between learned circuit blocks; this classifier family was introduced by Pérez-Salinas et al. [7]. Rotation dropout is also not introduced here. Scala et al. [8] proposed dropout for quantum neural networks as a regularization approach intended to improve generalization. Our question is different: whether masking trainable rotations can act as a dependable defense when the inference rule is attacked with knowledge of its randomness.

Adversarial robustness in QML. Adversarial vulnerability and defense evaluation for quantum learning models were studied before this work. Lu et al. [5] developed a quantum adversarial-learning setting; West et al. [10] reported a broader adversarial-robustness benchmark; and Wendlinger et al. [9] compared adversarial robustness for quantum and classical learning models. These studies establish the security setting and the relevance of adversarial training and attack evaluation. They do not test the rotation-dropout rule assessed here.

Randomized and noise-based QML defenses. Randomness as a defense idea in QML is also prior work. Du et al. [2] investigated quantum noise for protecting quantum classifiers. Huang and Zhang [4] tested a noise-layer QNN defense using quantum adversarial training as a baseline and stated that their tests assume adversaries know the randomization process. Gong et al. [3] studied randomized encodings for adversarial robustness. These mechanisms differ from masking learned rotation angles, but they rule out any broad claim that randomized or stochastic QML defenses, or informed evaluation of them, begin with this paper.

Novelty of this study. Our literature search found prior quantum dropout, prior QML adversarial-robustness work, and prior randomized or noise-based defenses, but no close report of the same trainable rotation-dropout evaluation matrix under ordinary PGD, adaptive EOT-PGD, transferred attacks, and adversarial-training comparisons. Therefore, our original contribution is a bounded defense-specific evaluation: it applies an existing trainable-rotation dropout mechanism to an

Table 1. Evaluation protocol. The required study uses exact statevector simulation; the four-qubit model is reported only as a confirmation check.

Component	Fixed specification
Data	Circle and Moons; two features in $[-1,1]^2$; 200/200/1000 train/validation/test samples; balanced attack subset of 256 test examples; five seeds.
Required QNNs	DR-QNN-1Q: one-qubit data re-uploading classifier. Angle-VQC-2Q: two-qubit angle-encoded variational classifier.
Training regimes	Standard (STD), quantum dropout (QD), adversarial training (AT), and adversarial training with quantum dropout (AT-QD).
Dropout	Trainable elementary rotations only; applied angle $m_j\theta_j$ with $m_j \sim \text{Bernoulli}(1-p^*)$; $p^* = 0.05$ for both required QNNs; encoders and CNOT gates are never masked.
Optimization	Adam; batch size 32; learning rates 0.05 (DR-QNN-1Q) and 0.01 (Angle-VQC-2Q); maximum epochs 100 (STD/QD) and 60 (AT/AT-QD).
Attacks	ℓ_∞ constraint; $\epsilon \in \{0.025, 0.05, 0.10, 0.20\}$; PGD20 with three starts; EOT-PGD20 with two starts and $K = 8$ mask samples per gradient estimate.
Stochastic testing	Monte Carlo inference uses $M = 16$ masks; EOT-PGD20 governs all security claims for stochastic modes.
Reporting	Clean accuracy, robust accuracy, attack success rate, NLL, Brier score, ECE, and QNode execution count.

attack-aware security test, reports that it does not provide a consistent adaptive robustness benefit in the fixed small-QNN protocol, and records the QNode execution cost of testing that stochastic inference rule fairly.

4 Models, Rotation Dropout, and Experimental Protocol

4.1 Study design and data

We test whether stochastic masking of trainable quantum rotations improves adversarial robustness after the attacker is allowed to account for the same randomness used at inference. The required study comprises two binary, two-feature tasks (Circle and Moons), two small QNN families, four training regimes, and five paired dataset seeds. Table 1 gives the fixed settings.

For each seed $s \in \{0, 1, 2, 3, 4\}$, each task contains 200 training, 200 validation, and 1000 clean-test examples. The Circle task draws $x = (x_1, x_2)$ uniformly from $[-1, 1]^2$ and sets

$$y = \mathbb{I}\left[\|x\|_2 < \sqrt{2/\pi}\right]. \tag{6}$$

The Moons task uses `make_moons` with noise 0.1. Its features are mapped to $[-1, 1]^2$ using a min–max transform fitted on the training partition only; transformed validation and test values are clipped to this domain. Both datasets use

stratified partitions. A fixed attack subset is selected from each test split by drawing exactly 128 samples per class without replacement using seed $70000 + s$, yielding 256 attacked examples per dataset and seed. We report clean metrics on all 1000 test examples and attack metrics on this fixed balanced subset.

4.2 Quantum probability models

A QNN maps an input x and trainable angles θ to a state prepared by a parameterized circuit U and then measures the readout wire. With dropout mask m, the binary class probabilities are

$$|\psi_{\theta,m}(x)\rangle = U_{\theta,m}(x)|0\rangle^{\otimes n}, \qquad p_{\theta,m}(c \mid x) = \langle\psi_{\theta,m}(x)|\Pi_c|\psi_{\theta,m}(x)\rangle, \tag{7}$$

where n is the number of wires and Π_c projects readout wire 0 onto class $c \in \{0, 1\}$. Both reported models return the two probabilities directly, and training minimizes negative log-likelihood with probability floor 10^{-8}.

Data-reuploading QNN (DR-QNN-1Q). Our one-wire model follows the repeated-encoding principle of data re-uploading [7]. It contains three layers and nine trainable rotation angles. In circuit order, each layer first applies $\mathrm{Rot}(x_1, x_2, 0)$, followed by trainable R_X, R_Y, and R_Z gates on wire 0:

$$U^{\mathrm{DR}}_{\theta,m}(x) = \prod_{\ell=1}^{3} \left[R_Z(m_{\ell,3}\theta_{\ell,3})R_Y(m_{\ell,2}\theta_{\ell,2})R_X(m_{\ell,1}\theta_{\ell,1})\,\mathrm{Rot}(x_1, x_2, 0)\right]. \tag{8}$$

This circuit repeatedly inserts the same two continuous features between trainable processing blocks.

Angle-encoded variational QNN (Angle-VQC-2Q). The two-wire model embeds x_1 and x_2 as X-axis angles once, then applies six trainable layers. Each layer contains R_X, R_Y, and R_Z on both wires followed by a $\mathrm{CNOT}(0, 1)$ coupling gate. It therefore has 36 trainable rotation angles in total ($6 \times 2 \times 3$). Both models use measurement probabilities on wire 0 as the binary output.

Optional four-wire confirmation. The optional Circle-only check uses a separate Angle-VQC-4Q model with two trainable layers and 24 trainable angles. It embeds $[x_1, x_2, x_1, x_2]$ on four wires and applies an R_X–R_Y–R_Z block followed by a CNOT ring in each layer. The attacker still perturbs only the original two input coordinates. This condition is reported as a limited confirmation check, not as evidence of a qubit-count trend.

4.3 Trainable rotation dropout and training regimes

We adapt quantum dropout from the QNN regularization setting of Scala et al. [8]. The dropout unit in this study is an elementary *trainable* rotation angle

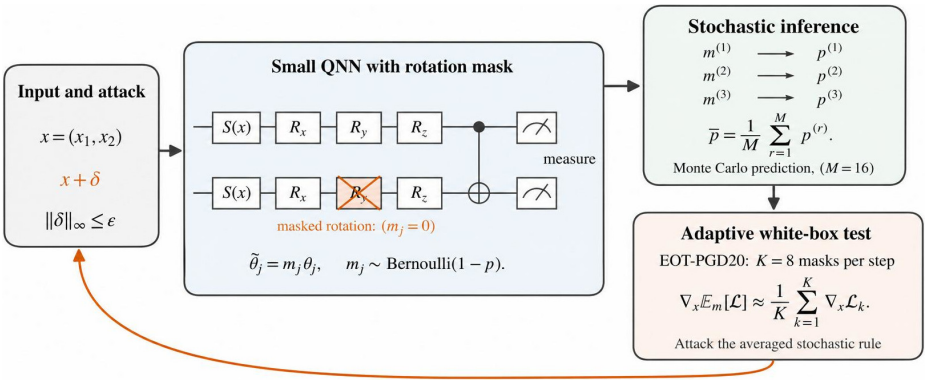

Fig. 1. Attack-aware evaluation of trainable rotation dropout. A bounded input perturbation is evaluated by a QNN whose trainable rotation gates may be masked. Stochastic prediction averages M mask realizations; the adaptive attack estimates gradients through K independently sampled masks at each update.

only. Encoding gates and CNOT coupling gates are never masked. For trainable angle θ_j, a binary mask is sampled as

$$m_j \sim \text{Bernoulli}(1-p), \qquad \widetilde{\theta}_j = m_j \theta_j, \tag{9}$$

so that $m_j = 0$ sets one trainable rotation angle to zero for that circuit evaluation. A pilot based only on seed-0 training and validation splits selected $p^* = 0.05$ for each required model family, and selected learning rates 0.05 for DR-QNN-1Q and 0.01 for Angle-VQC-2Q. These settings were frozen before final test evaluation.

We compare four regimes: standard clean training (STD), clean training with rotation dropout (QD), adversarial training without dropout (AT), and adversarial training with dropout (AT-QD). All models are optimized with Adam and batch size 32. STD and QD use at most 100 epochs; AT and AT-QD use at most 60 epochs. Within each dataset+model, all four regimes start from identical initialized parameters. QD and AT-QD sample a fresh mask for each minibatch loss evaluation. For AT-QD, one sampled mask is reused for construction of that minibatch's adversarial examples and for the outer parameter update.

For STD and QD, the retained checkpoint minimizes full-circuit validation negative log-likelihood, with higher validation accuracy and then earlier epoch as tie breakers. For AT and AT-QD, the retained checkpoint maximizes full-circuit PGD5 validation robust accuracy at $\epsilon = 0.10$, checked every five epochs; tie breakers are higher clean validation accuracy, lower clean validation negative log-likelihood, and earlier epoch. Neither Monte Carlo inference nor test outputs are used for checkpoint selection. At inference time, stochastic modes average probabilities over $M = 16$ independent rotation masks:

$$\overline{p}_\theta(c \mid x) = \frac{1}{M} \sum_{r=1}^{M} p_{\theta, m^{(r)}}(c \mid x), \qquad M = 16. \tag{10}$$

4.4 Threat model and attack-aware evaluation

The attacker performs untargeted evasion in the processed two-dimensional feature space. Given labeled input (x, y), feasible perturbations satisfy

$$\Delta_\epsilon(x) = \{\delta : \|\delta\|_\infty \le \epsilon,\ x + \delta \in [-1, 1]^2\}, \qquad \max_{\delta \in \Delta_\epsilon(x)} \mathcal{L}(q_\theta(\cdot \mid x + \delta), y)\,, \quad (11)$$

where $q_\theta = p_{\theta,\mathbf{1}}$ for deterministic full-circuit prediction and $q_\theta = \overline{p}_\theta$ for stochastic inference. Perturbation budgets are $\epsilon \in \{0.025, 0.05, 0.10, 0.20\}$ for the predeclared primary cell, Circle/DR-QNN-1Q; all cross-model comparisons use $\epsilon = 0.10$.

Following the robust-optimization attack convention of projected gradient descent (PGD) [6], ordinary PGD20 uses three random starts, 20 updates, and step size $\alpha = \epsilon/5$:

$$\delta_{t+1} = \mathit{\Pi}_{\Delta_\epsilon(x)}[\delta_t + \alpha\ \mathrm{sign}(\nabla_x \mathcal{L}(q_\theta(\cdot \mid x + \delta_t), y))]\,. \quad (12)$$

FGSM is retained as a one-step ordinary white-box diagnostic. Adversarial training uses PGD5 at $\epsilon = 0.10$, step size 0.025, and one zero-start construction per minibatch.

Because stochastic inference randomizes the gradient seen by an attacker, an ordinary attack against one fixed or all-active circuit is not sufficient for the security claim. Consistent with attack evaluation of randomized classifiers [1], adaptive EOT-PGD20 estimates the loss gradient by averaging over $K = 8$ independently sampled masks at each update:

$$\nabla_x\, \mathbb{E}_m[\mathcal{L}] \ \approx\ \frac{1}{K} \sum_{k=1}^{K} \nabla_x \mathcal{L}\big(p_{\theta,m^{(k)}}(\cdot \mid x), y\big)\,, \qquad K = 8. \quad (13)$$

EOT-PGD20 uses two random starts, 20 steps, and $\alpha = \epsilon/5$. It governs every robustness conclusion for QD or AT-QD under Monte Carlo inference.

A transfer check is reported separately. For each dataset and seed, a surrogate multilayer perceptron with widths 2–32–32–2, ReLU activations, Adam learning rate 0.01, batch size 32, and validation-loss checkpoint selection is trained on the same split. PGD20 perturbations are constructed on the surrogate and then evaluated on the QNN checkpoints. Transfer results serve as a black-box-style check and do not replace adaptive white-box evaluation.

4.5 Metrics, cost accounting, and simulation boundary

Let $\widehat{y}_i = \arg\max_c q_\theta(c \mid x_i)$ and $\widehat{y}_i^{\mathrm{adv}} = \arg\max_c q_\theta(c \mid x_i + \delta_i)$. We report clean accuracy, robust accuracy, and attack success rate conditioned on initially correct examples:

$$\mathrm{CA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\widehat{y}_i = y_i], \qquad \mathrm{RA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\widehat{y}_i^{\mathrm{adv}} = y_i], \quad (14)$$

$$\mathrm{ASR} = \frac{\sum_i \mathbb{I}[\widehat{y}_i = y_i]\mathbb{I}[\widehat{y}_i^{\mathrm{adv}} \ne y_i]}{\sum_i \mathbb{I}[\widehat{y}_i = y_i]}. \quad (15)$$

Additional recorded measures are macro-F1, negative log-likelihood, the two-class Brier score $N^{-1}\sum_i\sum_c(q_\theta(c \mid x_i) - \mathbb{I}[y_i = c])^2$, and expected calibration error with ten equal-width confidence bins. Metrics are aggregated by mean and standard deviation across five paired seeds; paired robust-accuracy differences use descriptive paired 95% t intervals.

The computational measure is the recorded number of QNode executions, supplemented by forward-pass equivalents and elapsed wall time. This matters for stochastic defense testing because Monte Carlo prediction and EOT gradient estimation repeat circuit evaluations. Results separate the required experiment scope from the optional four-wire check.

All reported QNN computations use PennyLane `default.qubit` with exact statevector simulation, `shots=None`, the PyTorch automatic-differentiation interface, and double-precision model inputs and parameters. The accepted execution environment used Python 3.11.15, PennyLane 0.45.0, PyTorch 2.12.0+cpu, NumPy 2.4.6, and scikit-learn 1.8.0 on CPU. The experiment does not report execution on quantum hardware or finite-sampling behavior. Although dataset files are not distributed with the paper, the generator rules, split sizes, scaling procedure, seed rules, model definitions, training rules, attacks, and metric conventions above fix an independent rerun of the study.

5 Results

5.1 Security reporting rule

The central question is whether stochastic quantum rotation dropout remains beneficial when the attack accounts for the same random masking used at inference. We therefore report deterministic methods under PGD20 and stochastic methods under adaptive EOT-PGD20 with Monte Carlo (MC) prediction. Unless stated otherwise, the primary comparison uses $\epsilon = 0.10$, five matched dataset seeds, and the fixed balanced attack subset of 256 test examples per seed. For stochastic modes, $M = 16$ masks are averaged at inference and $K = 8$ masks are used per EOT gradient estimate. The contrasts used below are

$$\Delta_{\mathrm{QD}} = \mathrm{RA}_{\mathrm{QD+MC+EOT}} - \mathrm{RA}_{\mathrm{STD+PGD}}, \tag{16}$$

$$\Delta_{\mathrm{ATQD}} = \mathrm{RA}_{\mathrm{AT\text{-}QD+MC+EOT}} - \mathrm{RA}_{\mathrm{AT+PGD}}. \tag{17}$$

where RA denotes robust accuracy. Positive values favor the method containing dropout; negative values favor its matched non-dropout comparison. Confidence intervals are descriptive paired 95% t intervals over five seeds.

5.2 Core adaptive-security result

Table 2 and Fig. 2 give the main result. Rotation dropout alone does not provide a consistent adaptive robustness benefit over standard training. In the predeclared primary cell, Circle / DR-QNN-1Q, the dropout method decreases robust accuracy

Table 2. Required results at $\epsilon = 0.10$. Entries are mean ± standard deviation over five seeds. Robust columns use PGD20 for STD and AT, and adaptive EOT-PGD20 with Monte Carlo inference for QD and AT-QD. Bold marks the highest robust-accuracy mean in each cell without a significance claim.

Cell	Metric	STD	QD	AT	AT-QD
Circle / DR-QNN-1Q	Clean	0.886 ± 0.026	0.876 ± 0.035	0.901 ± 0.026	0.892 ± 0.012
	Robust	**0.705 ± 0.022**	0.681 ± 0.026	0.702 ± 0.031	0.685 ± 0.013
Circle / Angle-VQC-2Q	Clean	0.617 ± 0.043	0.631 ± 0.064	0.706 ± 0.051	0.539 ± 0.090
	Robust	0.539 ± 0.027	0.546 ± 0.032	**0.587 ± 0.027**	0.463 ± 0.093
Moons / DR-QNN-1Q	Clean	0.887 ± 0.009	0.888 ± 0.009	0.892 ± 0.006	0.878 ± 0.011
	Robust	**0.834 ± 0.015**	0.831 ± 0.024	0.830 ± 0.028	0.828 ± 0.021
Moons / Angle-VQC-2Q	Clean	0.866 ± 0.010	0.860 ± 0.008	0.873 ± 0.006	0.866 ± 0.016
	Robust	0.812 ± 0.023	0.810 ± 0.022	**0.825 ± 0.020**	0.817 ± 0.024

Table 3. Paired robust-accuracy contrasts at $\epsilon = 0.10$ (left method minus matched baseline). Adaptive intervals are descriptive 95% paired t intervals over five seeds; transfer rows report paired means only.

A. Adaptive white-box evaluation: mean difference [95% interval]

Cell	Δ_{QD}	Δ_{ATQD}
Circle / DR-QNN-1Q	-0.023 [-0.033, -0.014]	-0.016 [-0.059, +0.026]
Circle / Angle-VQC-2Q	+0.007 [-0.030, +0.044]	-0.123 [-0.248, +0.001]
Moons / DR-QNN-1Q	-0.002 [-0.015, +0.010]	-0.002 [-0.019, +0.015]
Moons / Angle-VQC-2Q	-0.002 [-0.009, +0.006]	-0.008 [-0.028, +0.012]

B. Transfer evaluation: paired mean difference

Cell	Δ_{QD}	Δ_{ATQD}
Circle / DR-QNN-1Q	-0.016	-0.012
Circle / Angle-VQC-2Q	+0.007	-0.038
Moons / DR-QNN-1Q	+0.002	-0.002
Moons / Angle-VQC-2Q	-0.002	-0.003

Δ_{QD}: QD minus STD; Δ_{ATQD}: AT-QD minus AT under matched tests.

from 0.705 ± 0.022 to 0.681 ± 0.026. In the remaining required cells, the dropout-alone mean difference is small: it is positive only for Circle / Angle-VQC-2Q and remains close to zero for both Moons cells.

The combined-defense question is clearer. AT-QD with stochastic inference is below AT in mean robust accuracy in every required cell. The largest observed gap occurs for Circle / Angle-VQC-2Q, where AT reaches 0.587 ± 0.027 robust accuracy whereas AT-QD reaches 0.463 ± 0.093. Thus, in this protocol, adding rotation dropout does not improve the adversarially trained method and may reduce its utility in a model-dependent setting.

5.3 Paired contrasts and transferred attacks

Table 3 reports the paired contrasts that control the manuscript claims. For dropout alone, the primary Circle / DR-QNN-1Q contrast is $\Delta_{\mathrm{QD}} = -0.023$ with

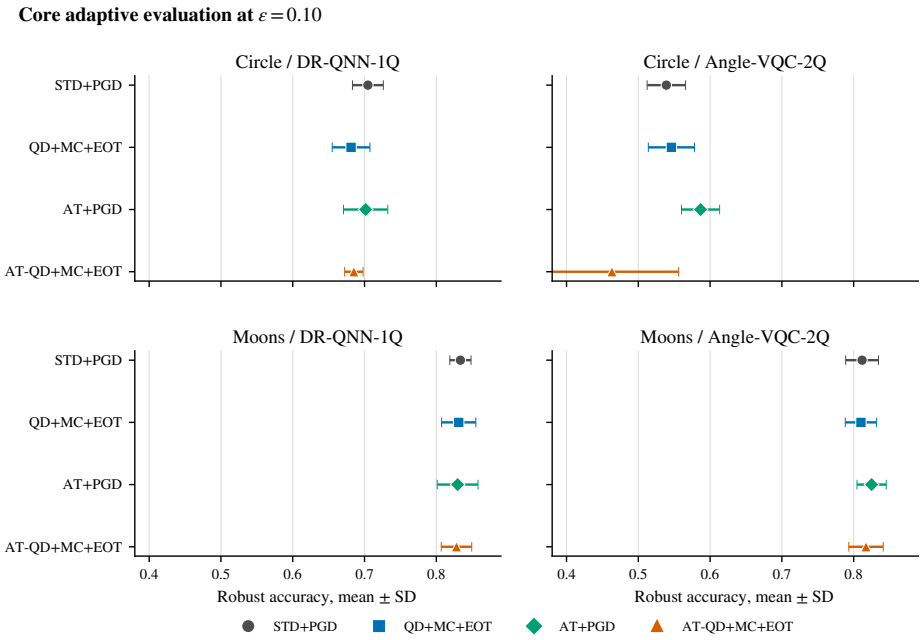

Fig. 2. Required robust-accuracy results at $\epsilon = 0.10$ under the reportable attack for each mode. Stochastic modes are evaluated by EOT-PGD20 with MC prediction. Error bars show standard deviation over five seeds.

interval $[-0.033, -0.014]$, and dropout is lower in all five paired seeds. The only positive required-cell mean for dropout alone is Circle / Angle-VQC-2Q, $+0.007$ with interval $[-0.030, +0.044]$; this is a limited observation, not evidence of a dependable defense.

For the combined method, all four required contrasts are negative. The Circle / Angle-VQC-2Q value is $\Delta_{\mathrm{ATQD}} = -0.123$ with interval $[-0.248, +0.001]$ and is lower in all five paired seeds. The interval slightly includes zero, so the result is stated as the largest observed mean reduction and a consistent paired-seed direction, not as an interval-separated effect.

The transfer check supports the same practical reading. In each required cell, AT-QD remains below AT in mean transferred-attack robust accuracy, with differences from -0.002 to -0.038. Dropout alone again lacks a consistent direction across cells. Transfer results therefore support, but do not replace, the adaptive white-box conclusion.

5.4 Perturbation budget and stochastic-evaluation stability

Fig. 3 examines the predeclared primary cell across all tested perturbation budgets. At $\epsilon = 0.025$, robust accuracy is 0.863 for STD, 0.845 for QD, 0.876 for AT, and 0.850 for AT-QD. At $\epsilon = 0.10$, the values are 0.705, 0.681, 0.702, and 0.685,

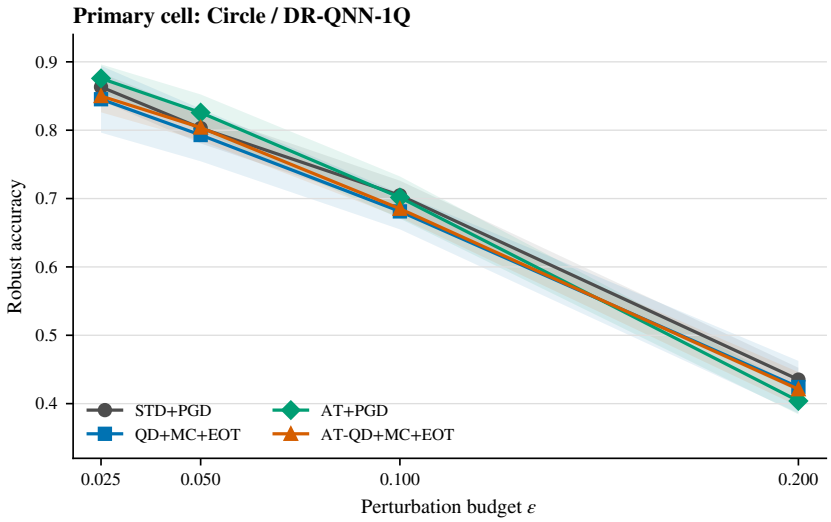

Fig. 3. Robust accuracy across perturbation budgets for the predeclared primary cell, Circle / DR-QNN-1Q. Bands show one standard deviation over five seeds.

respectively. At the strongest tested perturbation, $\epsilon = 0.20$, all four methods fall below 0.44 robust accuracy. Consequently, the primary-cell conclusion is not produced by a single isolated perturbation budget: dropout does not dominate the non-dropout comparisons across the tested range.

The predeclared stability checks support the chosen stochastic evaluation settings. In the primary QD cell at $\epsilon = 0.10$, varying the EOT gradient-mask count from $K = 1$ to $K = 16$ leaves mean robust accuracy at 0.68125 for every tested value. Varying MC inference from $M = 4$ to $M = 32$ produces mean robust accuracies from 0.680 to 0.682, while the single-mask setting is lower at 0.627. We therefore retain $K = 8$ and $M = 16$ as stable, predeclared settings.

5.5 Optional four-qubit confirmation check

The optional Circle / Angle-VQC-4Q check is presented separately in Table 4 and is not interpreted as a qubit-count trend. QD with MC and EOT has a small positive mean contrast over STD, $\Delta_{\mathrm{QD}} = +0.009$ with interval $[-0.036, +0.054]$. In contrast, AT-QD remains below AT in mean robust accuracy, $\Delta_{\mathrm{ATQD}} = -0.046$ with interval $[-0.159, +0.067]$. The optional check therefore does not reverse the required result: it gives, at most, a small uncertain dropout-alone observation while retaining a negative mean direction for the combined method.

5.6 Execution-count cost of fair stochastic testing

The attack-aware conclusion has a measurable computational cost. As reported in Table 4, EOT-PGD20 construction followed by MC prediction requires 352

Table 4. Additional four-qubit confirmation and execution-count accounting. The four-qubit result is a confirmation check only and cannot support a scaling claim.

A. Additional Circle / Angle-VQC-4Q result at $\epsilon = 0.10$		
Method	Clean acc.	Robust acc.
STD+PGD	0.855 ± 0.055	0.637 ± 0.033
QD+MC+EOT	0.846 ± 0.044	0.647 ± 0.023
AT+PGD	0.868 ± 0.046	0.657 ± 0.025
AT-QD+MC+EOT	0.774 ± 0.100	0.611 ± 0.069
Δ_{QD} adaptive	+0.009 [-0.036, +0.054]	
Δ_{ATQD} adaptive	-0.046 [-0.159, +0.067]	
B. QNode execution counts		
Scope		Executions
Required core training and adversarial evaluation		16,133,680
Additional four-qubit training and clean evaluation		1,523,400
Additional four-qubit adversarial evaluation		1,429,760
Total including four-qubit confirmation		19,086,840

QNode executions per attacked example, compared with 64 for PGD20 followed by deterministic prediction, a protocol-specific ratio of 5.5×. The required experiment matrix consumed 16,133,680 QNode executions; the optional four-qubit check adds 2,953,160, for 19,086,840 executions when both scopes are counted. This cost is scientifically useful: it quantifies the additional work required to test a randomized QNN prediction rule against an attacker that accounts for its randomness rather than accepting a potentially optimistic ordinary-attack result.

Result summary. Under the required adaptive evaluation, rotation dropout does not provide a consistent robust-accuracy benefit, and adding it to adversarial training does not improve mean robustness in any required cell. The transfer check and the optional four-qubit check do not alter that conclusion. The outcome is constructive for QML security: it identifies a plausible regularization-based defense that should not be credited with robustness without an attack-aware evaluation, and it provides a fixed cost baseline for testing later stochastic defenses under the same standard.

6 Discussion, Limitations, and Reproducibility

Security meaning of the negative result. Quantum rotation dropout is reasonable to test because quantum dropout was introduced as a QNN regularizer [8]. Security demands more: an attacker may optimize against the deployed prediction rule, including its random masks. This is the adaptive-testing principle for randomized classifiers [1]; in QML, noise-layer defense has likewise been studied while assuming that the attacker knows the randomization process [4]. Under that rule, rotation dropout does not give a consistent benefit over standard training, and its addition

to adversarial training gives lower mean robust accuracy than adversarial training alone in all four required cells. Thus, a plausible regularizer is not a verified defense merely because it changes gradients or performs well under a weaker attack. The negative result is useful security evidence: it prevents unsupported robustness credit for this mechanism under the tested protocol.

Research opportunity. The result is specific to trainable rotation dropout, not to every stochastic circuit defense. In the tested two-wire variational classifier, adversarial training without dropout improves mean robust accuracy over standard training for both datasets, whereas the dropout-augmented form adds no benefit. This pattern points toward defense-aware training objectives rather than independent random removal of trainable rotations. At the selected low dropout rate, EOT does not reveal a large hidden failure relative to ordinary stochastic evaluation; dropout instead lacks a dependable gain and can interfere with the stronger trained model. Later studies should predeclare and test expected masked-loss adversarial training, gate-role mask rules, and finite-sampling calibration objectives. These are new questions, not results of this study.

Limits of the evidence. All reported QNN calculations use exact-state CPU simulation, not a quantum processor or finite-sampling measurement. The required matrix includes two generated binary tasks with two features, one one-wire re-uploading classifier, one two-wire variational classifier, five paired seeds, and a selected low dropout rate. The four-wire Circle condition is additional confirmation only and does not establish a circuit-size pattern. We do not test images, high-dimensional inputs, deeper circuits, broad dropout-rate searches, physical noise, or device execution. The paired intervals are descriptive summaries over five fixed seeds, not broad statistical guarantees. Likewise, the $5.5\times$ result is a QNode execution-count ratio for the stated attack and inference rule, not a runtime or hardware-speed claim.

Reproducibility. Both tasks are generated from stated rules: an independent rerun can generate Circle and Moons for seeds $s \in \{0, 1, 2, 3, 4\}$; create stratified splits of 200 training, 200 validation, and 1000 test examples; and select the balanced attack subset with seed $70000 + s$ and 128 test examples per class. The Methods section fixes scaling, gates, parameter counts, mask scope, selected learning and dropout rates, paired initialization, checkpoint selection, PGD and EOT settings, the transfer surrogate, metrics, and QNode accounting. The executed stack was Python 3.11.15; PennyLane 0.45.0 with `default.qubit` and `shots=None`; PyTorch 2.12.0+cpu; NumPy 2.4.6; and scikit-learn 1.8.0. Computations used double-precision QNN values on CPU.

7 Conclusion

We tested whether quantum rotation dropout remains useful when attacked through its stochastic prediction rule. Across the required exact-simulation

matrix, it gives no consistent adaptive benefit, and adding it to adversarial training does not improve mean robust accuracy in any required comparison. Transfer and additional four-wire checks do not change that conclusion.

This negative result gives a practical security rule: randomized quantum classifiers deserve robustness credit only after their deployed stochastic rule is attacked directly. Under this protocol, adaptive stochastic testing requires 5.5× the QNode executions per attacked example of ordinary deterministic PGD evaluation. Future work should test attack-aware masked-loss training, gate-role masks, and finite-sampling settings under predeclared protocols.

Acknowledgments. Part of this work was funded by the National Science Foundation under grant CNS-2136961. The author thanks Dr. Gissella Bejarano Nicho for facilitating access to high-performance computing equipment. The author thanks the Rivas.AI Lab (`https://lab.rivas.ai`) for the support and helpful feedback throughout this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 274–283 (2018). `https://doi.org/10.48550/arXiv.1802.00420`
2. Du, Y., Hsieh, M.H., Liu, T.: Quantum noise protects quantum classifiers against adversaries. Physical Review Research **3**(2), 023153 (2021). `https://doi.org/10.1103/PhysRevResearch.3.023153`
3. Gong, W., Dong, Y., Li, W., et al.: Enhancing quantum adversarial robustness by randomized encodings. Physical Review Research **6**(2), 023020 (2024). `https://doi.org/10.1103/PhysRevResearch.6.023020`
4. Huang, C., Zhang, S.: Enhancing adversarial robustness of quantum neural networks by adding noise layers. New Journal of Physics **25**(8), 083019 (2023). `https://doi.org/10.1088/1367-2630/ace8b4`
5. Lu, S., Duan, L.M., Deng, D.L.: Quantum adversarial machine learning. Physical Review Research **2**(3) (2020). `https://doi.org/10.1103/PhysRevResearch.2.033212`
6. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018). `https://doi.org/10.48550/arXiv.1706.06083`
7. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., Latorre, J.I.: Data re-uploading for a universal quantum classifier. Quantum **4**, 226 (2020). `https://doi.org/10.22331/q-2020-02-06-226`
8. Scala, F., Ceschini, A., Panella, M., Gerace, D.: A general approach to dropout in quantum neural networks. Advanced Quantum Technologies **8**(12) (2023). `https://doi.org/10.1002/qute.202300220`

9. Wendlinger, M., Tscharke, K., Debus, P.: A comparative analysis of adversarial robustness for quantum and classical machine learning models. In: IEEE International Conference on Quantum Computing and Engineering. pp. 1447–1457 (2024). https://doi.org/10.1109/QCE60285.2024.00171
10. West, M.T., Erfani, S.M., Leckie, C., Sevior, M., Hollenberg, L.C.L., Usman, M.: Benchmarking adversarially robust quantum machine learning at scale. Physical Review Research **5**(2), 023186 (2023). https://doi.org/10.1103/PhysRevResearch.5.023186