IOP Publishing New J. Phys. 28 (2026) 014511 https://doi.org/10.1088/1367-2630/ae32a8

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

PAPER

Dual-regularized nonlinear quantum encoding for adversarial robustness in quantum machine learning

YaoChong Li[1,2], XinXin Deng[1,2,*], RuiQing Xu[3], WenShan Xu[1,2] and Ri-Gui Zhou[1,2]

[1] College of Information Engineering, Shanghai Maritime University, Shanghai, People's Republic of China
[2] Research Center of Intelligent Information Processing and Quantum Intelligent Computing, Shanghai Maritime University, Shanghai, People's Republic of China
[3] Faculty of Intelligence Technology, Shanghai Institute of Technology, Shanghai, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: dxxabc@outlook.com

Keywords: quantum neural networks, adversarial attacks, robustness, nonlinear transformations

OPEN ACCESS

RECEIVED
10 September 2025

REVISED
11 December 2025

ACCEPTED FOR PUBLICATION
2 January 2026

PUBLISHED
21 January 2026

Abstract

Quantum machine learning (QML) models are highly vulnerable to adversarial attacks, severely limiting their deployment in security-sensitive applications. To enhance their robustness, this study proposes a novel method, termed nonlinear quantum encoding with dual regularisation (NQE-DR), which combines nonlinear quantum encoding with dual regularisation. NQE-DR integrates classical nonlinear transformations, adaptive parameter optimisation, and a multi-qubit cross-coupling architecture. Our dual regularisation incorporates parameter and gradient regularisation, providing a theoretical framework to guarantee robustness under adversarial attacks and noisy conditions. Experimental results demonstrate that NQE-DR significantly outperforms mainstream encoding methods, exhibiting a notable improvement in classification accuracy and enhanced stability under FGSM attacks and noise. This study advances adversarial quantum machine learning by seamlessly integrating nonlinear encoding with a comprehensive dual regularisation strategy, offering a potent security enhancement for practical QML deployment.

1. Introduction

Quantum computing [1, 2], leveraging superposition, entanglement, and parallelism, offers a novel paradigm for overcoming computational bottlenecks in complex problems. Quantum machine learning (QML) [3–5], a frontier field at the intersection of quantum computing and artificial intelligence, utilises parametrised quantum circuits (PQCs) to process classical data, demonstrating potential advantages in tasks such as image classification [6] and natural language processing [7]. However, as QML extends into security-sensitive domains, including financial risk control [8, 9] and medical diagnosis [10, 11], concerns over adversarial robustness have intensified [12, 13]. Attackers can introduce minor perturbations to input data, significantly dropping the prediction accuracy of quantum models and severely restricting the practical deployment of QML [14].

Current research on quantum adversarial defence primarily explores adversarial training [15, 16], quantum noise [17, 18], quantum circuit architecture [19, 20] and regularisation strategies [21, 22], yet these methods exhibit some limitations. For example, parameter regularisation [23] may enhance generalisation by constraining parameter complexity, but often fails to adequately mitigate sensitivity to input perturbations. In contrast, gradient regularisation [24] can reduce the magnitude of model gradients to smooth decision boundaries, but this may compromise the model's learning capacity. Moreover, most of these approaches adopt a single-perspective approach, lacking a comprehensive defence framework that integrates both data encoding and model training. To address these limitations, this study proposes a novel unified approach combining nonlinear quantum encoding with dual regularisation (NQE-DR). This framework balanced expressibility, global correlation, and robustness by integrating classical nonlinear transformations, fully connected quantum entanglement, and dual regularisation. The proposed

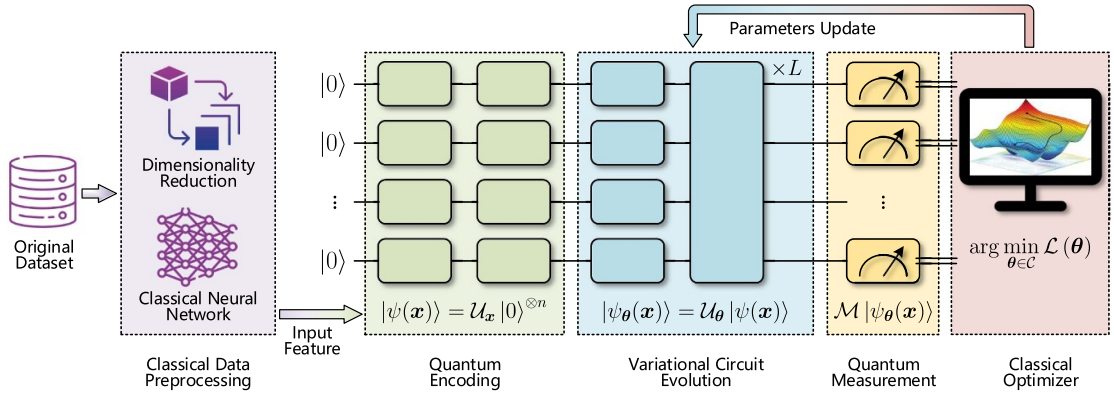


Figure 1. Schematic illustration of a generic quantum neural network pipeline. The workflow begins with classical data preprocessing (e.g. dimensionality reduction or a classical neural network), followed by quantum encoding (such as amplitude, angle, or hardware-efficient methods), unitary evolution of quantum states through variational circuits, and quantum measurement. The measured outcomes are processed by a classical optimiser, ultimately yielding classical predictive outputs. In this work, the classical preprocessing step is implemented as a procedure comprising data normalization, dimensionality reduction, and a dedicated nonlinear transformation, which are detailed in section 3.1.

method provides a new theoretical foundation for enhancing QML robustness and practical implications for deploying quantum technologies in security-sensitive scenarios.

The innovative contributions of this study are as follows.

- We introduce the NQE-DR model, which synergistically combines nonlinear feature interaction through classical transformations, fully connected quantum entanglement to establish global qubit correlations, and dual regularisation (parameter and gradient) to constrain model complexity and perturbation sensitivity. This integrated design enriches feature representation, controls the model's Lipschitz bound, and enhances adversarial robustness.
- We perform comprehensive experiments on multiple benchmark datasets, including synthetic data (Circle, Gaussian Mixture), real-world clinical data (Breast Cancer), and a large-scale image dataset (MNIST), to evaluate the proposed method across diverse data complexities and dimensions.
- We provide extensive comparative analysis demonstrating that NQE-DR significantly outperforms not only mainstream encoding methods but also advanced approaches such as problem-driven feature maps (e.g. ZZFeatureMap) and data re-uploading architectures under both noise interference and Fast Gradient Sign Method (FGSM) attacks.

The rest of this paper is organised as follows. Section 2 reviews relevant concepts of related technology. Section 3 details the overall architecture of NQE-DR, including classical nonlinear transformation, quantum encoding layer, entanglement enhancement layer, adaptive measurement layer, and the dual regularisation mechanism. Section 4 presents comparative experiments, adversarial attack tests, scalability analysis and ablation experiments to validate NQE-DR's expressiveness and robustness. Section 5 concludes the method's results, limitations, and future research directions for quantum encoding in large-scale quantum computing.

2. Preliminaries

2.1. Quantum neural network (QNN) and encoding methods

QNNs have emerged as a cutting-edge paradigm bridging quantum computing and machine learning [25, 26]. By harnessing unique quantum characteristics, including superposition, entanglement, and gate-based operations, they provide new capabilities for data processing and analysis.

As shown in figure 1, a typical QNN workflow involves four sequential operations: data preprocessing, quantum encoding, variational circuit evolution, and quantum measurement. Its architecture consists of three core components: a quantum data encoder, a PQC, and a classical optimiser. The data encoder maps classical features into quantum states; the PQC applies trainable quantum gates whose parameters are iteratively tuned to capture data patterns; and the classical optimiser updates these parameters by minimising a loss function computed from the measurement outcomes.

Among these components, the quantum data encoder constitutes one of the essential modules, as it specifies how classical information is transformed into quantum states for subsequent processing. The commonly used quantum encoding methods are summarized as follows:

- Amplitude encoding (AmE) [27] encodes the elements of a classical data vector $\boldsymbol{x}$ into the amplitudes of a quantum state. Specifically, a classical vector of dimension 2^n can be represented as an n-qubit quantum state, given by

$$|\psi(\boldsymbol{x})\rangle = \frac{1}{||\boldsymbol{x}||}\sum_{i=0}^{2^n-1} x_i |i\rangle, \tag{1}$$

  where x_i denotes the components of the data vector and $|i\rangle$ represents the computational basis states. This approach achieves a high degree of data compression; however, it typically requires deep and complex quantum circuits and is highly sensitive to noise.
- Angle encoding (AE) [27], also referred to as phase or rotational encoding, maps classical data onto the Bloch sphere of single qubits through rotation gates. For an n-dimensional classical vector $\boldsymbol{x}$, each component x_i is encoded as a rotation angle, such that

$$|\psi(\boldsymbol{x})\rangle = \bigotimes_{i=1}^{n} R_y(x_i)|0\rangle, \tag{2}$$

  where $R_y(x_i)$ denotes a rotation about the Y-axis by angle x_i. This method is straightforward to implement on quantum hardware; nevertheless, it exhibits low encoding density and often requires a large number of qubits to represent high-dimensional data.
- Hardware-efficient encoding (HEE) [28, 29] is a quintessential example of a hardware-aware feature map, as its design philosophy is intrinsically guided by the physical constraints and native capabilities of quantum processing units. This approach explicitly exploits the native gate structures and connectivity constraints of specific quantum devices to minimize circuit depth, reduce operational errors, and enhance implementation fidelity. It typically integrates single-qubit rotation gates (e.g. R_X, R_Y, R_Z) native to the hardware with available two-qubit entangling gates, thereby embedding classical features into quantum states through a hardware-optimized circuit. Although the specific implementation of HEE depends on the underlying hardware architecture, its overarching objective is to design encoding circuits that maximize efficiency and robustness by leveraging the intrinsic characteristics of the hardware platform, making it a foundational strategy for deploying QML models on near-term quantum devices.

2.2. Advanced quantum encoding paradigms

Beyond the standard encoding methods, recent research has evolved towards more specialized paradigms that are either driven by specific problem structures or architectural innovations.

- **Problem-driven feature maps** refer to encoding strategies whose design is explicitly guided by the inductive biases of the target task or desired quantum state properties. A prominent example is the **ZZFeatureMap** [30], which employs a layered structure of data embedding and entanglement. For an input data vector $\boldsymbol{x} \in \mathbb{R}^n$, a single layer of the ZZFeatureMap can be described by the unitary operation:

$$U_{\mathrm{ZZFM}}(\boldsymbol{x}) = \left(\bigotimes_{i=1}^{n} R_Z(x_i)\right) \cdot \left(\bigotimes_{i=1}^{n} R_Y(x_i)\right) \cdot \left(\bigotimes_{i,j\in E} ZZ_{i,j}(x_i x_j)\right) \tag{3}$$

  where E defines the set of qubit pairs for entangling interactions, often a linear chain or all-to-all connectivity. The entangling gate $ZZ_{i,j}(\theta) = \mathrm{e}^{-\mathrm{i}\theta(Z_i\otimes Z_j)/2}$ creates correlations between features. This design aims to create highly entangled, complex feature spaces that are particularly suitable for capturing intricate correlations in data, making it a principled approach for enhancing model expressivity.
- **Data Re-uploading** [21, 31] represents a fundamentally different architectural paradigm that blurs the traditional boundary between encoding and variational circuits. Instead of a single data encoding step followed by a parametrised transformation, a data re-uploading model with L layers is defined by a sequence of alternating encoding and variational circuits:

$$|\psi(\boldsymbol{x})\rangle = U^{(L)}_{\boldsymbol{\theta}_L} U_{\mathrm{enc}}(\boldsymbol{x}) \cdots U^{(2)}_{\boldsymbol{\theta}_2} U_{\mathrm{enc}}(\boldsymbol{x}) \cdot U^{(1)}_{\boldsymbol{\theta}_1} U_{\mathrm{enc}}(\boldsymbol{x}) |0\rangle^{\otimes n} \tag{4}$$

Here, $U_{\text{enc}}(\boldsymbol{x})$ is a data-encoding unitary (e.g. a layer of $R_Y(x_i)$ gates), and $U_{\boldsymbol{\theta}_l}^{(l)}$ is the lth trainable variational layer. This deep encoding strategy significantly enhances the model's nonlinear representation capacity without increasing the number of qubits, effectively creating a quantum analogue of deep classical networks by allowing the data to be processed multiple times at different circuit depths.

2.3. Adversarial quantum learning and regularisation strategies

With the rapid advancement of QNNs, their security vulnerabilities have become increasingly evident. Adversarial quantum learning investigates the robustness of QML models against deliberately crafted perturbations. Similar to the classical setting [32, 33], an adversarial attack introduces a subtle and often imperceptible perturbation $\boldsymbol{\delta}$ to an input $\boldsymbol{x}$, resulting in an adversarial example $\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \boldsymbol{\delta}$ that misleads the QNN into producing an incorrect prediction. Even slight perturbations can significantly alter the quantum state in the Hilbert space, thereby inducing large variations in prediction outcomes. A representative attack method is the FGSM, which derives adversarial perturbations $\boldsymbol{\delta} = \epsilon \cdot \text{sgn}(\nabla_{\boldsymbol{x}} \mathcal{L}(f(\boldsymbol{x}), y))$ based on the gradient of the model's loss function with respect to the input data. The resulting perturbation is subsequently added to the original input to generate the adversarial example.

Over the past five years, the concept of adversarial quantum learning has been introduced and has attracted increasing attention. In response, a variety of defence strategies have been proposed, which can be broadly classified into the following categories:

- Adversarial training [15, 16] represents a straightforward yet effective defence approach. By incorporating adversarial examples into the training set together with clean samples, the model learns to maintain predictive accuracy under adversarial perturbations, thereby improving its robustness in practice.
- Quantum noise [17, 18], while often viewed as a limiting factor for near-term quantum devices, can also be exploited as a defensive mechanism. The controlled introduction of stochastic fluctuations during the inference process may attenuate adversarial perturbations by reducing their relative effect on the final prediction.
- Quantum circuit architecture [19, 20] has also been identified as a determinant of adversarial robustness. Circuit properties such as expressibility, entanglement capacity, and the choice of parametrised gates shape the representation of data within the Hilbert space, and consequently influence the vulnerability of QNNs to adversarial manipulation.
- Regularisation strategies [21, 22] aim to constrain model behaviour through modifications of the loss function. Parameter regularisation introduces penalties on trainable parameters to mitigate overfitting and enhance generalisation, whereas gradient regularisation suppresses large gradients with respect to the input, thereby lowering sensitivity to small perturbations and smoothing decision boundaries.

Among the various defence strategies, regularisation-based methods have emerged as an important line of research aimed at constraining model behaviour and improving robustness. Existing regularisation schemes often concentrate on single-dimensional constraints. For example, parameter regularisation facilitates generalisation but only marginally reduces the model's sensitivity to input perturbations. Similarly, gradient regularisation smooths decision boundaries, but it may compromise the model's learning capacity and accuracy. Furthermore, few studies integrate these regularisation approaches with data encoding defences to create a more comprehensive defence framework. Building upon this idea, this study proposes a dual regularisation scheme that combines parameter and gradient regularisation. The former limits the range of the model's trainable parameters to prevent overfitting and over-sensitivity to parameter perturbations, and the latter penalises the norm of the loss function's gradient with respect to the input to control the model's Lipschitz bound and reduce sensitivity to minor input perturbations. This combined strategy offers a novel and efficient way to improve robustness in QNN models against adversarial attacks.

2.4. Lipschitz constants and adversarial robustness evaluation

A key factor in assessing adversarial robustness is the model's sensitivity to input perturbations, which can be theoretically quantified using the *Lipschitz constant* [34–36]. Formally, a function $f: \mathcal{X} \to \mathbb{R}^C$ is said to be *L-Lipschitz* with respect to the p-norm if there exists a constant $L \geqslant 0$ such that, for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|_p \leqslant L \|\boldsymbol{x} - \boldsymbol{x}'\|_p. \tag{5}$$

Here, $\|\cdot\|_p$ denotes the p-norm. This inequality implies that the maximum change in the model's output is linearly bounded by the size of the input perturbation. In adversarial scenarios,

when $\|\boldsymbol{\delta}\|_p \leqslant \epsilon$, one has

$$\|f(\boldsymbol{x}+\boldsymbol{\delta}) - f(\boldsymbol{x})\|_p \leqslant L\epsilon, \tag{6}$$

indicating that a smaller Lipschitz constant yields tighter control of the output under perturbations and thus improves adversarial robustness. In the above, $\boldsymbol{\delta}$ represents the adversarial perturbation, ϵ the perturbation budget, and L the Lipschitz constant.

For differentiable functions, the Lipschitz constant can be characterised via the Jacobian of f. The *local Lipschitz constant* around $\boldsymbol{x}$ is given by the operator norm of the Jacobian:

$$L(\boldsymbol{x}) = \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|_{p\to p}, \tag{7}$$

where the operator norm is defined as

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|_{p\to p} = \sup_{\|\boldsymbol{u}\|_p=1} \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\,\boldsymbol{u}\|_p. \tag{8}$$

The *global Lipschitz constant* is then obtained as $L = \sup_{\boldsymbol{x}\in\mathcal{X}} L(\boldsymbol{x})$. Here, $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \in \mathbb{R}^{C\times d}$ denotes the Jacobian matrix of f with respect to the input, and $\|\cdot\|_{p\to p}$ is the operator norm induced by the p-norm.

From the above, it follows that a smaller Lipschitz constant implies that adversarial perturbations must be of larger magnitude to alter the model's prediction, thereby enhancing robustness. However, in practice, directly computing or tightly controlling the global Lipschitz constant of deep or quantum neural networks is generally intractable. A widely adopted approach is to employ *regularisation*, which acts either explicitly or implicitly to constrain the Lipschitz constant. For instance, parameter regularisation (e.g. weight decay or spectral normalisation) bounds the operator norm of the weight matrices, thereby limiting the global Lipschitz constant, while gradient regularisation penalises the Jacobian norm to reduce local Lipschitz values. In both cases, regularisation restricts the model's sensitivity to input variations, establishing a principled connection between Lipschitz continuity and adversarial robustness.

3. Design of NQE-DR scheme

This section presents the architecture and core modules of the NQE-DR scheme, as illustrated in figure 2. The input data $\boldsymbol{x}$ is first processed by a nonlinear classical transformation $\mathcal{T}$ to enrich feature interactions; the transformed features are then mapped into quantum states through the encoding module $\mathcal{E}$; global correlations are introduced via the entanglement block $\mathcal{U}$ent; task-specific learning is performed by the trainable variational block $\mathcal{U}$trainable; finally, measurement $\mathcal{M}$ extracts classical outputs. The framework incorporates dual regularisation to constrain the Lipschitz bound and enhance robustness.

3.1. Overall framework

The primary objective of NQE-DR is to construct a quantum encoding pipeline that achieves both high expressiveness and robustness. Its overall framework is expressed as:

$$\text{NQE-DR}(\boldsymbol{x}) = \mathcal{M} \circ \mathcal{U}_{\text{trainable}} \circ \mathcal{U}_{\text{ent}} \circ \mathcal{E} \circ \mathcal{T}(\boldsymbol{x}) \tag{9}$$

where $\mathcal{T}$ denotes the classical nonlinear transformation module (sigmoid and squaring operation); $\mathcal{E}$ represents the quantum encoding module (mapping classical features to quantum states); $\mathcal{U}_{\text{ent}}$ signifies the fully connected entanglement module (CNOT network); $\mathcal{U}_{\text{trainable}}$ refers to the adaptive phase rotation module (trainable R_Y gates); $\mathcal{M}$ corresponds to the joint measurement module (extracting classical outputs). The symbol $\circ$ denotes the composition of operations, meaning that each module operation is executed sequentially from right to left.

The classical nonlinear transformation is essential for linking high-dimensional inputs with quantum encoding in NQE-DR, mapping raw features into a nonlinear feature space suitable for quantum gate operations through two-stage transformations.

Inspired by the preprocessing employed in [37], for an input vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^n$, feature normalisation and nonlinear enhancement are first performed through sigmoid transformation:

$$\boldsymbol{x}_{\text{sig}} = \sigma(\boldsymbol{x}) = \left[\frac{1}{1+\mathrm{e}^{-x_1}}, \frac{1}{1+\mathrm{e}^{-x_2}}, \ldots, \frac{1}{1+\mathrm{e}^{-x_n}}\right] \tag{10}$$

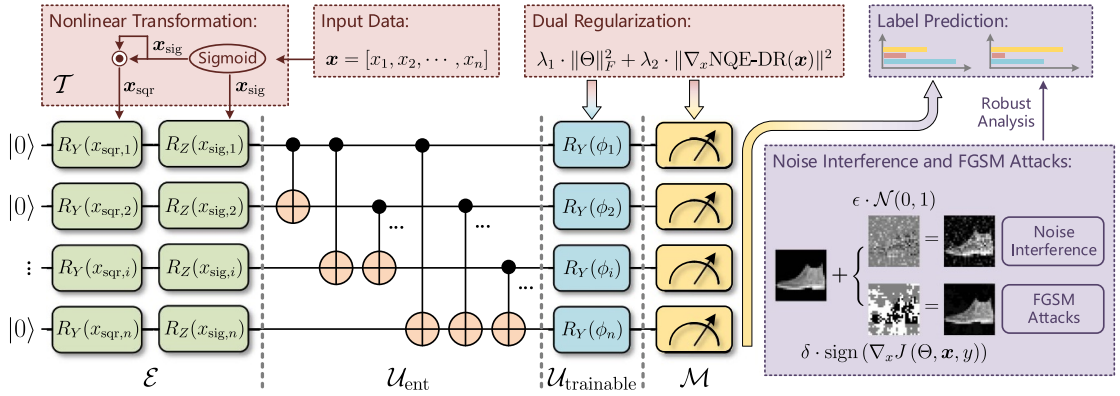


Figure 2. Schematic of the NQE-DR framework. Input $\boldsymbol{x}$ undergoes nonlinear transformation $\mathcal{T}$, quantum encoding $\mathcal{E}$, entanglement $\mathcal{U}$ent, trainable evolution $\mathcal{U}$trainable, and measurement $\mathcal{M}$ for classical output. Dual regularisation constrains the Lipschitz bound and improves robustness.

where $\sigma(x) \in (0,1)$. This transformation compresses the input dynamic range to prevent quantum gate rotation angle saturation from extreme values and introduces soft nonlinearity to enrich feature gradients for subsequent quantum encoding. To capture second-order feature correlations, an element-wise squaring operation is applied to $\boldsymbol{x}_{\text{sig}}$:

$$\boldsymbol{x}_{\text{sqr}} = \boldsymbol{x}_{\text{sig}} \odot \boldsymbol{x}_{\text{sig}} = \left[x_{\text{sig},1}^2, x_{\text{sig},2}^2, \ldots, x_{\text{sig},n}^2\right] \tag{11}$$

where $\odot$ denotes the Hadamard product. The squaring operation amplifies local sensitivity by expanding feature differences. For instance, when $x_{\text{sig},i} \ll x_{\text{sig},j}$, the gap between $x_{\text{sqr},i}$ and $x_{\text{sqr},j}$ further expands, making quantum encoding more capable of distinguishing subtle feature differences. After $\mathcal{T}$ processing, the original input $\boldsymbol{x}$ is transformed into two sets of nonlinear features $\{\boldsymbol{x}_{\text{sig}}, \boldsymbol{x}_{\text{sqr}}\}$, providing dual-path feature inputs for quantum encoding.

The quantum encoding module converts classical nonlinear features into quantum states by mapping through rotation angle parametrisation. NQE-DR uses the tensor product state of n qubits as the initial state:

$$|\psi_0\rangle = |0\rangle^{\otimes n} = |0\rangle \otimes |0\rangle \otimes \cdots \otimes |0\rangle \tag{12}$$

For the ith qubit, we adopt an R_Y–R_Z dual-gate encoding scheme, which shares similarities with circuit 1 in [38] . The rotation angles of the R_Z and R_Y gates are determined by $\boldsymbol{x}$sig and $\boldsymbol{x}$sqr, respectively, such that $\theta_{Z,i} = x_{\text{sig},i}$ and $\theta_{Y,i} = x_{\text{sqr},i}$. The dual-gate operation is mathematically expressed as:

$$U_i = R_Y(\theta_{Y,i}) R_Z(\theta_{Z,i}) = R_Y\left(x_{\text{sqr},i}\right) R_Z\left(x_{\text{sig},i}\right) \tag{13}$$

where the matrix forms of R_Y and R_Z are:

$$R_Y(\theta) = \begin{pmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{pmatrix}, \quad R_Z(\theta) = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2}. \end{pmatrix}$$

The input encoded state is generated through parallel encoding of n qubits:

$$|\psi_{\text{in}}\rangle = \mathcal{E}\left(\boldsymbol{x}_{\text{sqr}}\right) = \bigotimes_{i=1}^{n} U_i |\psi_0\rangle = \bigotimes_{i=1}^{n} R_Y\left(x_{\text{sqr},i}\right) R_Z\left(x_{\text{sig},i}\right) |0\rangle^{\otimes n}. \tag{14}$$

A notable feature of this quantum state is that each qubit's phase is directly modulated by the classical feature of nonlinear transformations. Additionally, the cross effect (non-commutativity) of R_Y and R_Z further introduces quantum-level nonlinearity, enriching the feature basis for subsequent entanglement enhancement. The entanglement enhancement module strengthens global qubit correlations through two-stage operations: fully connected entanglement and adaptive phase rotation, corresponding to the PQC Ansatz design. To establish global qubit correlations, NQE-DR employs a fully connected CNOT network to enhance entanglement, motivated by [39]:

$$|\psi_{\text{ent}}\rangle = \mathcal{U}_{\text{ent}} |\psi_{\text{in}}\rangle = \left(\bigotimes_{1 \leqslant i < j \leqslant n} \text{CNOT}_{i,j} \right) |\psi_{\text{in}}\rangle \tag{15}$$

where $\mathrm{CNOT}_{i,j}$ denotes the CNOT gate with the ith qubit as the control and the jth as the target, whose action is:

$$\mathrm{CNOT}_{i,j}|a\rangle_i \otimes |b\rangle_j = |a\rangle_i \otimes |a \oplus b\rangle_j \quad (a, b \in \{0,1\}).$$

Compared with local entanglement approaches (e.g. between adjacent qubits), the fully connected CNOT network offers $O(n^2)$ pairwise qubit correlations, surpassing the $O(n)$ correlations of local schemes, and quantum states that contain interactions among all feature pairs, providing support for capturing global patterns. To enhance model trainability, phase modulation with trainable R_Y gates is applied to the entangled state:

$$|\psi_{\text{output}}\rangle = \mathcal{U}_{\text{trainable}}|\psi_{\text{ent}}\rangle = \bigotimes_{i=1}^{n} R_Y(\phi_i)\,|\psi_{\text{ent}}\rangle \tag{16}$$

where ϕ_i denote trainable parameters, forming the parameter matrix $\Theta = [\phi_1, \phi_2, \ldots, \phi_n]^T \in \mathbb{R}^n$. The parameter ϕ_i adjusts qubit phase offsets, making the quantum state more likely to fit the task loss function. Their initial values are set as $\phi_i \sim \mathcal{U}(0, \pi)$ to avoid initial phase saturation. The measurement module converts quantum states into classical outputs. NQE-DR designs a joint measurement operator for the selective extraction of quantum features. NQE-DR adopts a structured measurement operator:

$$\mathcal{M} = \sigma_Z^{(0)} \otimes \sigma_X^{(1)} \otimes \mathbb{I}^{(2)} \otimes \cdots \otimes \mathbb{I}^{(n-1)} \tag{17}$$

where $\sigma_Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $\sigma_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ represents Pauli operators; $\sigma_Z^{(0)}$ denotes performing Z-measurement on the 0th qubit, and $\sigma_X^{(1)}$ signifies performing X-measurement on the 1st qubit; the remaining qubits (2nd to n–1th) maintain the identity operator $\mathbb{I}$ and do not participate in measurement to reduce information loss. The final output of NQE-DR is the expectation value of the measurement operator on the quantum state:

$$\text{NQE-DR}(\boldsymbol{x}) = \langle\psi_{\text{output}}|\mathcal{M}|\psi_{\text{output}}\rangle \tag{18}$$

This expectation value represents the projection average of the quantum state under the basis defined by $\mathcal{M}$, expressed mathematically as:

$$\text{NQE-DR}(\boldsymbol{x}) = \mathrm{Tr}\left(\rho_{\text{output}}\mathcal{M}\right) \quad \left(\rho_{\text{output}} = |\psi_{\text{output}}\rangle\langle\psi_{\text{output}}|\right)$$

where ρ_{output} signifies the density matrix of the quantum state. For classification, the output is converted into a probability distribution through the softmax function:

$$p(c|\boldsymbol{x}) = \frac{\exp\left(\text{NQE-DR}_c(\boldsymbol{x})\right)}{\sum_{c'=1}^{C}\exp\left(\text{NQE-DR}_{c'}(\boldsymbol{x})\right)}$$

where $\text{NQE-DR}_c(\boldsymbol{x})$ denotes the NQE-DR output for class c, obtained through multiple sets of measurement operators.

3.2. Dual regularisation mechanism

To control the model's Lipschitz bound and improve robustness, NQE-DR employs dual constraints of parameter and gradient regularisation. For the trainable parameter matrix Θ, a Frobenius norm constraint is adopted:

$$\mathcal{R}_1 = \lambda_1 \cdot \|\Theta\|_F{}^2 = \lambda_1 \cdot \sum_{i=1}^{n}\phi_i^2 \tag{19}$$

where $\lambda_1 > 0$ denotes the regularisation strength. This constraint (1) suppresses excessive growth of ϕ_i to prevent excessive sensitivity of quantum state phases to parameter perturbations; and (2) mitigate overfitting by limiting parameter complexity, theoretically confining the value range of Θ to $\|\Theta\|_F \leqslant \Theta_{\max}$, where $\Theta_{\max}$ negatively correlates with λ_1.

For the gradient of NQE-DR output with respect to the input, an L_2 norm constraint is adopted:

$$\mathcal{R}_2 = \lambda_2 \cdot \|\nabla_{\boldsymbol{x}}\text{NQE-DR}(\boldsymbol{x})\|^2 = \lambda_2 \cdot \sum_{i=1}^{n}\left(\frac{\partial\text{NQE-DR}(\boldsymbol{x})}{\partial x_i}\right)^2 \tag{20}$$

where $\lambda_2 > 0$ denotes the regularisation strength. This constraint directly controls the Lipschitz bound of the model:

$$L = \sup_{\boldsymbol{x}} \|\nabla_{\boldsymbol{x}} \text{NQE-DR}(\boldsymbol{x})\| \leqslant \sqrt{\frac{\mathcal{R}_2}{\lambda_2}} \leqslant G_{\max}$$

where $G_{\max}$ represents the upper bound of the gradient norm, reducing the model's sensitivity to input perturbations. The total loss function of NQE-DR integrates task loss and regularisation terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \cdot \|\Theta\|_F{}^2 + \lambda_2 \cdot \|\nabla_{\boldsymbol{x}} \text{NQE-DR}(\boldsymbol{x})\|^2 \tag{21}$$

where $\mathcal{L}_{\text{task}}$ denotes the specific task loss (e.g. cross-entropy loss for classification tasks). Quantum parameters Θ and regularisation strengths λ_1 and λ_2 are optimised simultaneously through backpropagation for expressiveness and robustness. This dual mechanism ensures that the model remains both functionally expressive and resistant to perturbations. By concurrently limiting parameter growth and output sensitivity, NQE-DR achieves a favourable trade-off between accuracy and stability, making it suitable for deployment in noisy or adversarial environments.

4. Numerical experiments

This section verifies the effectiveness of the NQE-DR encoding scheme through systematic experiments, including performance comparison, robustness evaluation, Lipschitz boundary computation, and ablation studies analysing the contributions of core modules. To comprehensively evaluate the performance of NQE-DR, we first compare it with mainstream quantum encoding methods on multiple classical benchmark datasets (Circle, Gaussian Mixture, Breast Cancer) to validate its fundamental effectiveness. Subsequently, to further investigate the model's scalability and robustness in more complex, high-dimensional scenarios and respond to the need for comparison with advanced quantum encoding paradigms, we design an extended experiment: an in-depth comparison of NQE-DR with advanced architectures such as problem-driven feature maps (ZZFeatureMap) and data re-uploading (Data Re-uploading) on the large-scale image dataset MNIST. This two-stage evaluation strategy from simple to complex ensures the comprehensiveness of the verification and the reliability of the conclusions.

4.1. Experimental setup

The details of the benchmark datasets are as follows:

- **Circle**: A circle with a radius of $\sqrt{2/\pi}$ is drawn within the range $X = [-1, +1] \times [-1, +1]$. Data points inside and outside the circle are labelled $y = +1$ and $y = 0$, respectively.
- **Gaussian Mixture**: A synthetic 10-dimensional Gaussian mixture dataset with three modes, comprising 1000 samples (700 training, 300 testing), is generated to verify the fitting ability for nonlinear distributions.
- **Breast Cancer (UCI)**: A 30-dimensional clinical dataset with features such as tumour radius and texture, containing 569 samples (400 training, 169 testing) of 357 benign, 212 malignant data. It is used to verify real-world robustness.

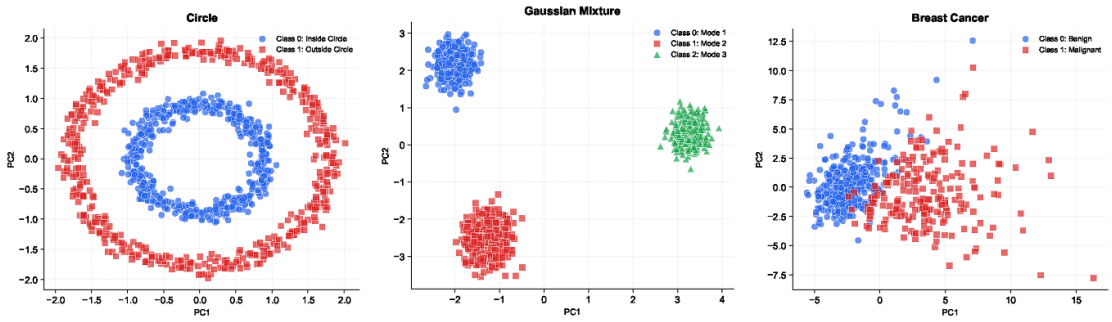
As visualized in figure 3, all datasets are standardised with features normalised to $[0, 1]$ to align with matching quantum rotation gate angles, and labels binarised (benign/inside-circle/Mode 1 $\rightarrow$ 0; malignant/outside-circle/Modes 2–3 $\rightarrow$ 1). The high-dimensional Breast Cancer dataset undergoes dimensionality reduction and correlation processing.

Comparison is conducted with the following recent quantum encoding schemes: **AmE**: maps classical features directly to quantum state amplitudes via unitary transformation, enabling efficient compression of high-dimensional classical data. **HEE**: combines local entanglement (e.g. nearest-neighbour CNOT gates) with trainable single-qubit rotation gates, widely adopted for practical implementations. **AE**: uses classical features directly as rotation angles for single-qubit gates.

All schemes employ the same PQCs framework.

Experimental parameters: The Adam optimiser [40] is used with a learning rate of $\eta = 0.1$ and 200 training epochs. Regularisation parameters $\lambda_1 = 0.01$ and $\lambda_2 = 0.2$ are selected via 5-fold cross-validation. Experiments are conducted using the PennyLane[4] framework [41], which is an open-source

[4] https://pennylane.ai/.



(a) Circle Dataset (b) Gaussian Mixture Dataset (c) Breast Cancer Dataset

Figure 3. PCA projections of the three datasets used in the experiments. From left to right: (a) The *Circle* dataset (two-class distribution with samples inside/outside a circle of radius $\sqrt{2/\pi}$); (b) The *Gaussian Mixture* dataset (three-modal high-dimensional data with 10 features, clustered into three distinct modes); (c) The *Breast Cancer* dataset (real-world clinical data with 30 features, distinguishing benign vs. malignant tumors). Color-coded markers represent different classes, and PCA retains the primary structural characteristics of the original data.

Table 1. Comparison of classification accuracy of four quantum encoding schemes: NQE-DR, HEE, AEE, and AmE on circle, Gaussian mixture, and Breast Cancer datasets. Higher accuracy indicates better adaptability and expressiveness of the encoding scheme to the dataset. In the table, the best-performing accuracy value for each dataset is highlighted in bold.

Dataset	NQE-DR	HEE	AE	AmE
Circle	**0.978**	0.945	0.930	0.880
Gaussian mixture	**0.985**	0.970	0.960	0.930
Breast Cancer	**0.958**	0.945	0.948	0.922

library for quantum machine learning, quantum chemistry, and quantum computing. It provides a simple and easy-to-use interface that supports multiple quantum simulators and hardware backends, facilitating the construction, training, and simulation of quantum circuits. Meanwhile, dependency management utilises Poetry[5] to ensure consistency and reproducibility of the experimental environment, including dataset selection and preprocessing. Additionally, Dask[6] is a flexible parallel computing library that can decompose large-scale computing tasks into multiple small tasks and execute them in parallel on multiple processors or nodes [42], significantly reducing the runtime of the experiment.

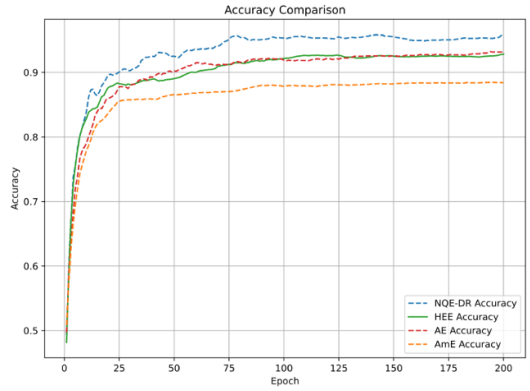
4.2. Experimental results and analysis

Table 1 presents the original accuracy of different encoding schemes on the three datasets. On all three datasets, NQE-DR consistently outperforms other schemes, achieving accuracies of 0.978 and 0.985 on the Circle and Gaussian Mixture datasets, respectively, demonstrating strong generalisation to different feature distributions. This superior performance stems from the synergistic effects of its core components: the nonlinear encoding module effectively captures high-order feature interactions, the fully connected entanglement network establishes global qubit correlations, and the dual regularisation balances expressiveness and stability. In the complex and high-dimensional Breast Cancer dataset, the gap between NQE-DR (0.958) and AE (0.948) narrows, but NQE-DR still maintains a notable lead.

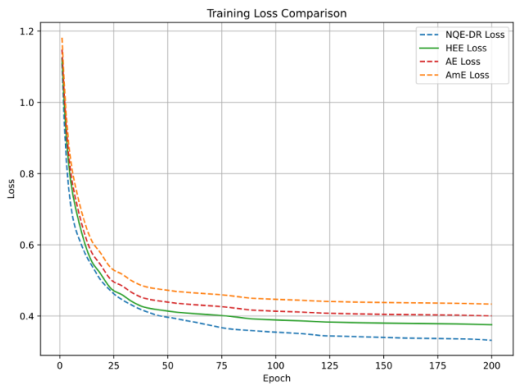
The accuracies of various encoding schemes on the circle classification problem are illustrated in figure 4(a), where NQE-DR consistently demonstrates superior performance. Specifically, NQE-DR achieves an accuracy of 97.8% on the circle dataset, surpassing all other baselines and verifying its capacity to capture high-order feature interactions essential for nonlinear boundary fitting. HEE ranks second with 94.5% accuracy, but its reliance on local entanglement (nearest-neighbour CNOT gates) restricts its ability to build global correlations, leading to inferior generalization compared with NQE-DR. In contrast, AmE achieves 88.0%, limited by its fixed amplitude mapping that lacks trainable flexibility. AE obtains 93.0%, capturing only linear correlations and failing to adapt to quadratic relationships, thereby underperforming relative to NQE-DR.

Similar advantages of NQE-DR are also observed in the Gaussian mixture and breast cancer datasets, as presented in figures 5(a) and 6(a). In both benchmarks, NQE-DR consistently exceeds baseline

[5] https://python-poetry.org/.
[6] www.dask.org/.

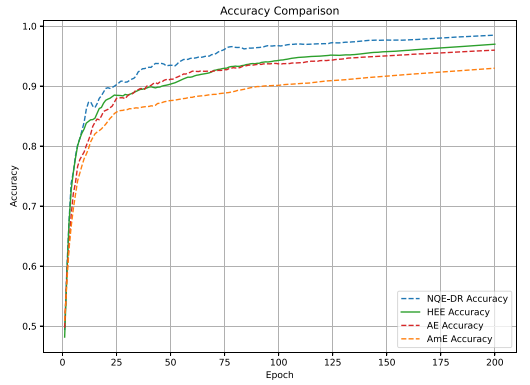


(a) Accuracy on Circle dataset

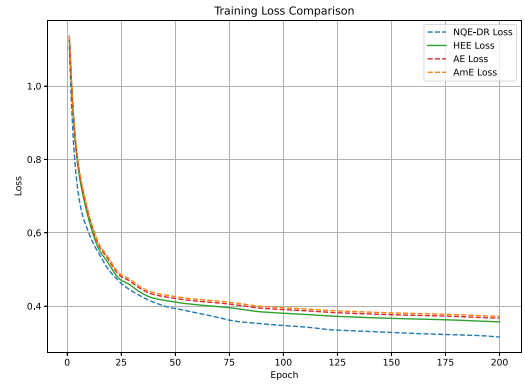


(b) Loss function on Circle dataset

Figure 4. Performance comparison between NQE-DR and baseline encoding methods on the Circle dataset: (a) accuracy evolution and (b) loss function behaviour.



(a) Accuracy on Gaussian Mixture dataset



(b) Loss function on Gaussian Mixture dataset

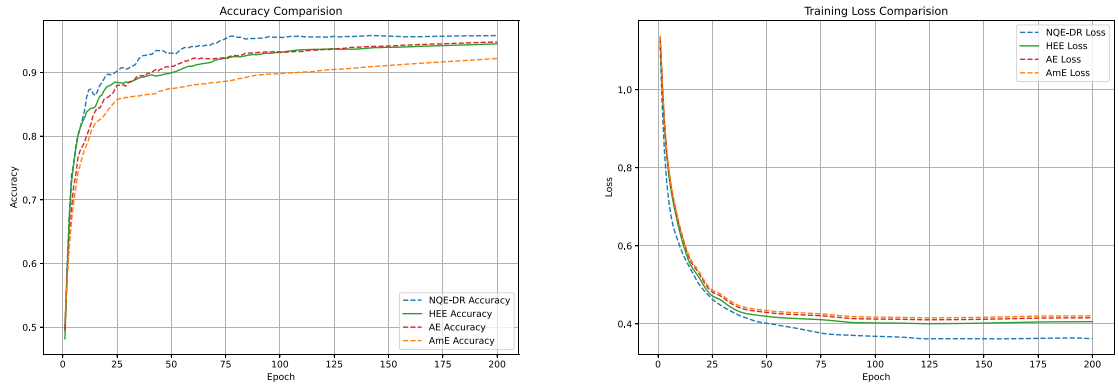
Figure 5. Performance comparison between NQE-DR and baseline encoding methods on the Gaussian Mixture dataset: (a) accuracy evolution and (b) loss function behaviour.

encodings by margins of 2%–5% in precision, confirming its robustness across heterogeneous data distributions.

From the perspective of training dynamics, the loss function analysis (figures 4(b)– 6(b)) further highlights the effectiveness of NQE-DR. On all three datasets, the proposed encoding not only converges more stably but also achieves a lower asymptotic loss compared with HEE, AmE, and AE. These results collectively demonstrate that NQE-DR is capable of both enhancing predictive accuracy and improving optimization efficiency across diverse classification tasks.

To evaluate NQE-DR's robustness relative to other encoding methods, experiments are conducted under varying noise intensities and FGSM-generated adversarial samples. For noise interference, we introduce a uniform random perturbation to the input features, where each component is independently sampled from , with controlling the noise intensity. The perturbed input is then clipped to the valid feature range before being fed into the quantum encoder. Table 2 summarises the performance of different quantum encoding schemes under this noise interference and FGSM attacks.

In scenarios involving noise interference and FGSM attacks across all three datasets, the accuracy of NQE-DR is generally higher than that of HEE, AE, and AmE. For example, on the Breast Cancer dataset at noise intensity 1.0, NQE-DR achieves accuracies of 0.553 and 0.528 under noise interference and FGSM attacks, respectively, exceeding those of other methods. This finding indicates that NQE-DR exhibits stronger robustness in complex interference environments, as its design incorporates classical nonlinear transformations, fully connected entanglement, and dual regularisation effectively, which collectively enhances resistance to interference. As the noise intensity increases from 0 to 1.0, the accuracy of all encoding schemes shows a downward trend. However, the decline in NQE-DR's accuracy is relatively smaller. For instance, on the Circle dataset, NQE-DR's accuracy decreases from 0.978 to 0.573, whereas AmE's falls from 0.880 to 0.454. This discrepancy highlights NQE-DR's superior resistance to



(a) Accuracy on Breast Cancer dataset (b) Loss function on Breast Cancer dataset

Figure 6. Performance comparison between NQE-DR and baseline encoding methods on the Breast Cancer dataset: (a) accuracy evolution and (b) loss function behaviour.

Table 2. The accuracy performance of four quantum encoding schemes: NQE-DR, HEE, AE, and AmE on circle, Gaussian mixture, and Breast Cancer dataset under varying levels of noise intensities and FGSM attacks. It compares their performance on both original (Clean) and perturbed data, highlighting the robustness differences among various encoding schemes. In the results, the best-performing accuracy value (for each dataset and condition) is highlighted in bold.

		Noise levels ϵ				FGSM attacks δ			
Dataset	Proportion	NQE-DR	HEE	AE	AmE	NQE-DR	HEE	AE	AmE
Circle	0.1	**0.938**	0.875	0.890	0.855	**0.945**	0.932	0.902	0.875
	0.2	**0.860**	0.825	0.841	0.831	**0.875**	0.900	0.874	0.843
	0.3	**0.821**	0.743	0.770	0.761	**0.842**	0.840	0.843	0.795
	0.4	**0.783**	0.731	0.711	0.660	**0.795**	0.795	0.794	0.754
	0.5	**0.764**	0.700	0.690	0.651	**0.754**	0.752	0.766	0.711
	0.6	**0.681**	0.669	0.599	0.612	**0.694**	0.683	0.690	0.656
	0.7	**0.652**	0.630	0.574	0.558	**0.665**	0.641	0.653	0.601
	0.8	**0.630**	0.589	0.564	0.529	**0.632**	0.594	0.583	0.563
	0.9	**0.591**	0.562	0.528	0.481	**0.587**	0.564	0.545	0.512
	1.0	**0.573**	0.505	0.501	0.454	**0.535**	0.511	0.492	0.462
Gaussian mixture	0.1	**0.931**	0.928	0.919	0.911	**0.920**	0.935	0.920	0.910
	0.2	**0.895**	0.877	0.874	0.854	**0.890**	0.885	0.821	0.832
	0.3	**0.845**	0.842	0.845	0.821	**0.851**	0.843	0.710	0.721
	0.4	**0.809**	0.785	0.775	0.743	**0.793**	0.711	0.658	0.620
	0.5	**0.763**	0.754	0.728	0.682	**0.754**	0.670	0.610	0.601
	0.6	**0.732**	0.712	0.705	0.621	**0.711**	0.659	0.531	0.572
	0.7	**0.658**	0.632	0.653	0.572	**0.692**	0.610	0.498	0.548
	0.8	**0.619**	0.592	0.585	0.510	**0.650**	0.582	0.464	0.439
	0.9	**0.545**	0.521	0.512	0.456	**0.621**	0.531	0.431	0.381
	1.0	**0.508**	0.482	0.477	0.398	**0.609**	0.490	0.375	0.334
Breast Cancer	0.1	**0.931**	0.918	0.921	0.895	**0.929**	0.915	0.903	0.882
	0.2	**0.894**	0.863	0.875	0.832	**0.890**	0.855	0.849	0.838
	0.3	**0.863**	0.829	0.834	0.789	**0.858**	0.819	0.807	0.782
	0.4	**0.821**	0.787	0.792	0.710	**0.797**	0.759	0.745	0.727
	0.5	**0.785**	0.729	0.719	0.652	**0.743**	0.719	0.686	0.665
	0.6	**0.731**	0.685	0.645	0.602	**0.698**	0.672	0.654	0.605
	0.7	**0.686**	0.649	0.603	0.565	**0.659**	0.642	0.630	0.558
	0.8	**0.653**	0.601	0.574	0.531	**0.626**	0.600	0.591	0.505
	0.9	**0.618**	0.561	0.532	0.478	**0.592**	0.549	0.532	0.463
	1.0	**0.553**	0.502	0.473	0.403	**0.528**	0.495	0.435	0.393

noise perturbations. Under FGSM attacks, NQE-DR also demonstrates better stability. At an attack intensity of 1.0 on the Gaussian Mixture dataset, NQE-DR yields 0.609 accuracy, compared to 0.490, 0.375, and 0.334 for HEE, AE, and AmE, respectively. NQE-DR maintains a relatively high accuracy even under strong attacks. These results verify that dual regularisation effectively controls the Lipschitz bound and reduces sensitivity to adversarial perturbations. NQE-DR reveals robust advantages in both

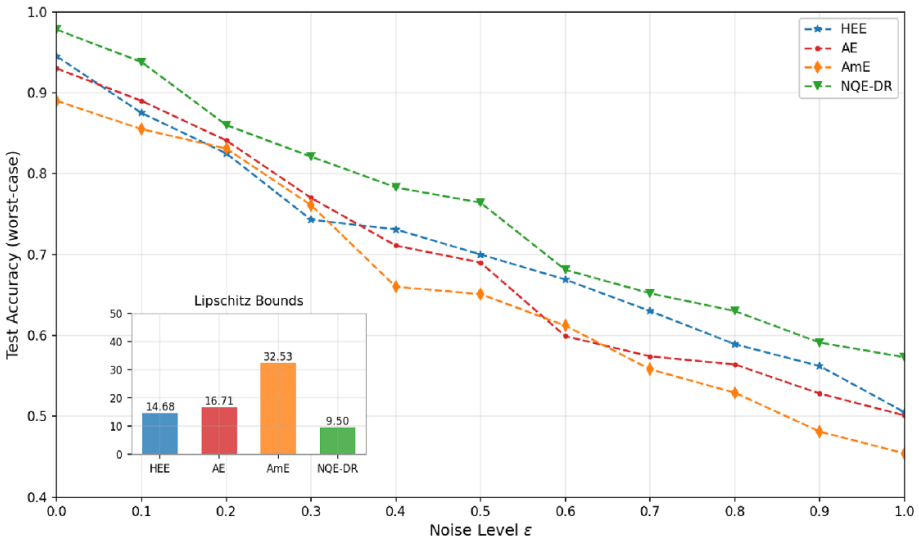


Figure 7. Robustness performance and Lipschitz bounds of different quantum encoding schemes under varying noise levels (ϵ). NQE-DR demonstrates consistently high robustness while maintaining the lowest Lipschitz value.

low-dimensional, simple-structured datasets (e.g. Circle) and high-dimensional, real-world medical datasets (e.g. Breast Cancer). This observation underscores the encoding structure's adaptability to data with different feature distributions and dimensions, validating its generalisation ability. Its fully connected entanglement network captures cross-dimensional feature correlations across datasets, classical nonlinear transformations model diverse nonlinear patterns, and dual regularisation ensures consistent robustness across scenarios.

Furthermore, figure 7 illustrates the relationship between the *Lipschitz bound* and robustness across different quantum encoding models under varying noisy environments ($\epsilon = 0.0$–1.0). The Lipschitz bound for each model was empirically estimated by averaging the local Lipschitz constants, , which were computed exactly for 50 validation samples using automatic differentiation. The results indicate that NQE-DR maintains high robustness across all noise levels and achieves the lowest *Lipschitz* value (9.5), which aligns with our theoretical expectations. This low Lipschitz bound stems from dual regularisation, where parameter regularisation limits rotation angle fluctuations, and gradient regularisation suppresses the sensitivity of quantum state outputs to input changes. In contrast, AmE yields the highest *Lipschitz* value, likely due to its fixed encoding structure with no trainable parameters to adjust for perturbation resistance, a critical factor influencing the *Lipschitz* property.

4.3. Scalability analysis

As described in the theoretical section,the Lipschitz bound plays a crucial role in characterizing model robustness through the input gradient norm $\|\nabla_{\boldsymbol{x}}\text{NQE-DR}(\boldsymbol{x})\|$. However, an important theoretical concern is that gradients in quantum circuits (including input gradients) may suffer from the *barren plateau* phenomenon as the system size increases, potentially rendering this regularization term ineffective.

To empirically validate the effectiveness of our method within the scale of current interest for the Noisy intermediate-scale quantum (NISQ) era and to explore its limits, we systematically analyzed how the Lipschitz value–calculated based on the input gradient norm–varies with quantum system size. We conducted measurements for different *qubit counts* $n \in [4, 6, 8, 10, 12]$ and *circuit depths* $d \in [3, 4, 5, 6, 8]$. For each (n, d) configuration,we estimated the Lipschitz value from multiple random input and parameter samples, reporting the average to ensure statistical reliability.

The results are shown in the 3D surface plot in figure 8. Key observations include:

- **Effectiveness at NISQ Scales:** In the parameter region with qubit count $n \leqslant 12$ and circuit depth $d \leqslant 8$ (which covers and exceeds the scales of the main experiments in this work),the Lipschitz value remains substantial. This provides direct empirical evidence for the effectiveness of Lipschitz regularization in our core experiments, confirming that it does not vanish due to barren plateaus at this scale.
- **Asymptotic decay trend:** As theoretically expected, we observe an overall decay in the Lipschitz value with increasing qubit count and circuit depth. This highlights the necessity of integrating our method

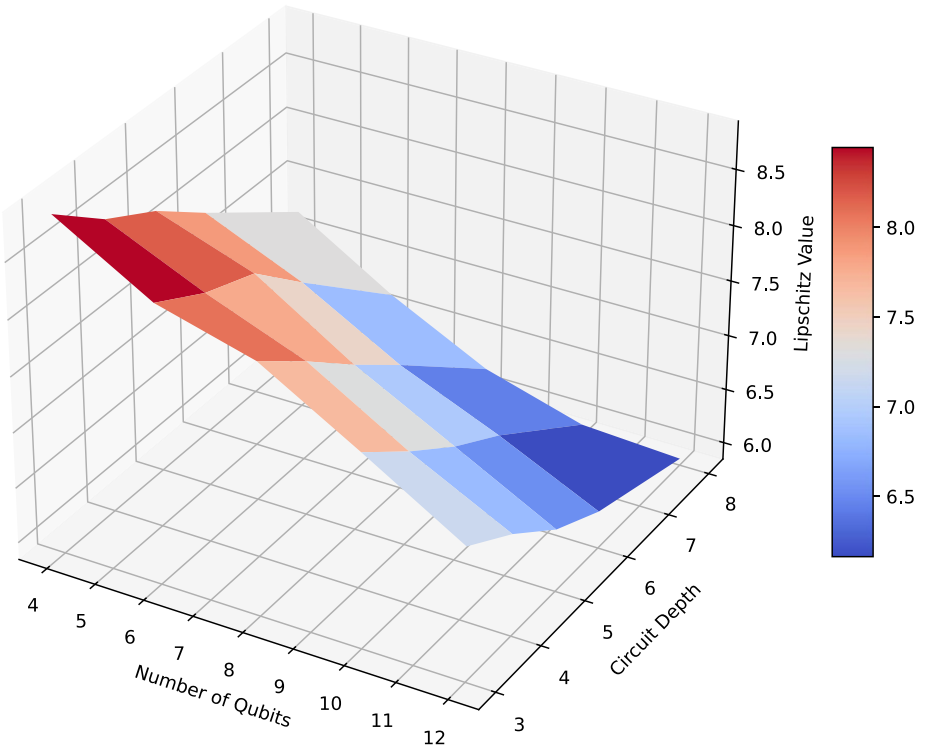


Figure 8. 3D surface plot of the Lipschitz value as a function of qubit count and circuit depth. The vertical axis represents the Lipschitz value estimated from input gradient norms, while the horizontal axes correspond to the system size parameters. The results reveal a controllable Lipschitz range at NISQ-relevant scales and highlight the gradual decay with increasing qubit count and circuit depth, consistent with theoretical expectations.

with more advanced barren plateau mitigation techniques when moving toward larger-scale quantum computing.

- **Positive role of architecture:** The specific design of our method, including the use of shallow-depth circuits and local measurements on a small number of qubits, aligns with the theory by Cerezo *et al* [43]. Our experimental results are consistent with this theoretical expectation.

Future work will systematically extend this analysis to deeper and larger quantum circuits to investigate the asymptotic scaling behaviour of the Lipschitz term and to develop adaptive layer-wise regularization strategies for mitigating gradient concentration effects at larger system scales.

To establish the generalizability and scalability of our proposed NQE-DR framework under realistic conditions, we conducted extensive evaluations on the MNIST dataset , a cornerstone benchmark in machine learning research. We focus on the binary classification task (digits 0 vs. 1) comprising 12 665 training and 2115 testing samples of 28 $\times$ 28 greyscale images. This configuration presents a significantly more challenging high-dimensional learning problem (784 features) compared to previous synthetic datasets, while maintaining computational tractability for rigorous quantum circuit simulations.

To address the critical challenge of mapping high-dimensional classical data to limited quantum resources, we employ a principled dimensionality reduction strategy. All images are flattened into 784-dimensional vectors, normalized to the [0,1] range, and subsequently projected to a 4-dimensional feature space using Principal component analysis (PCA). This approach represents a well-justified trade-off between maintaining essential data structure and accommodating NISQ-era hardware constraints, while ensuring all compared methods operate under identical preprocessing conditions for fair evaluation.

We establish a comprehensive benchmarking framework comparing NQE-DR against five strategically selected quantum encoding paradigms, representing the current state-of-the-art:

- **Standard encodings**:AmE and AE, HEE, oserving as baseline methods for classical-to-quantum information mapping.
- **ZZFeatureMap**: Employing structured entanglement through alternating R_Y/R_Z rotations and ZZ interactions to create physically motivated feature correlations.
- **Advanced hybrid architecture**: Data Re-uploading (DRU), implementing deep quantum networks through sequential data re-encoding and variational processing layers.

Table 3. Comparative analysis of quantum encoding schemes on MNIST under adversarial conditions. Performance metrics report classification accuracy across varying noise intensities (ϵ) and FGSM attack strengths (δ). The optimal accuracy value (for each combination of noise intensity and attack strength) is highlighted in bold.

		Noise ϵ				FGSM δ			
Encoding scheme	Clean	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
NQE-DR	**0.969**	**0.935**	**0.818**	**0.749**	**0.638**	**0.939**	**0.841**	**0.747**	**0.672**
DRU	0.969	0.922	0.801	0.736	0.624	0.929	0.823	0.739	0.643
ZZF	0.961	0.906	0.778	0.721	0.607	0.918	0.809	0.722	0.628
HEE	0.952	0.887	0.740	0.695	0.627	0.907	0.792	0.711	0.649
AE	0.945	0.871	0.765	0.686	0.571	0.898	0.762	0.685	0.639
AmE	0.929	0.851	0.757	0.647	0.553	0.872	0.750	0.707	0.597

The systematic comparison in table 3 reveals several crucial insights. First, NQE-DR demonstrates consistent superiority across all tested scenarios, achieving the highest accuracy in both pristine and adversarial environments. While DRU exhibits competitive performance in clean conditions (Accuracy = 0.969), its robustness degrades more significantly under strong adversarial perturbations ($\delta \geqslant 0.5$), highlighting the critical limitation of relying solely on architectural expressivity without explicit robustness constraints.

Furthermore, the performance hierarchy observed suggests a fundamental trade-off between encoding sophistication and adversarial robustness. Notably, the performance gap between NQE-DR and other methods widens systematically with increasing perturbation strength, providing compelling evidence for the necessity of integrated nonlinear encoding and dual regularization mechanisms. These results firmly establish that our approach not only maintains its advantages when scaling to complex, high-dimensional datasets but also delivers unmatched robustness against state-of-the-art quantum encoding strategies.

4.4. Ablation experiments

Ablation studies are conducted on the Circle dataset to quantitatively evaluate the individual contributions of the core modules within the NQE-DR framework. Four ablated variants are compared against the complete model: NQE-DR-$\mathcal{T}$ omits the classical nonlinear transformation ($\mathcal{T}$), directly encoding raw features without the sigmoid and squaring operations. NQE-DR-$\mathcal{U}_{\text{ent}}$ replaces the fully connected entanglement structure with a local entanglement scheme, where CNOT gates are restricted to adjacent qubits only. NQE-DR-$\mathcal{U}_{\text{layered}}$ substitutes the fully connected entanglement with a layered hardware-efficient ansatz, which employs nearest-neighbour entangling gates in a sequential layer-by-layer structure [45], representing a distinct and prevalent architectural paradigm. NQE-DR-$\mathcal{R}$ removes the dual regularisation ($\mathcal{R}$) by setting $\lambda_1 = \lambda_2 = 0$, thus training the model solely with the task loss.

The experimental results, summarized in table 4

Under different noise levels (0.1, 0.3, 0.5, 0.7) and FGSM adversarial attacks, the complete NQE-DR model achieves the highest accuracy and performance. At noise level and an FGSM attack intensity of 0.5, the full NQE-DR model attains accuracies of 0.764 and 0.754, respectively, while NQE-DR-$\mathcal{T}$ drops to 0.680 and 0.650. This decline reflects the variant's inability to capture complex nonlinear relationships in the data and high-order feature interactions, highlighting the notable impact of $\mathcal{T}$ on the integrity of feature expression. The variants with altered entanglement structures, NQE-DR-$\mathcal{U}_{\text{ent}}$ and NQE-DR-$\mathcal{U}_{\text{layered}}$, exhibit a moderate but clear performance drop relative to the full model. At a noise level of 0.5, their accuracies are 0.734 and 0.745, respectively. This confirms that while the fully connected entanglement provides the most expressive global correlations, alternative ansätze like local or layered entanglement still retain a substantial degree of robustness when coupled with the other components of NQE-DR. Notably, the layered ansatz ($\mathcal{U}$layered), being more structured and hardware-feasible than the local one ($\mathcal{U}$ent), bridges part of the performance gap to the full model, highlighting a favourable trade-off. Finally, the removal of the dual regularisation in NQE-DR-$\mathcal{R}$ leads to a notable decrease in robustness, with accuracy falling to 0.680 under an FGSM attack intensity of 0.5. This result directly attributes the model's stability to the regularisation mechanism, which effectively constrains the Lipschitz bound and suppresses excessive output fluctuations in the presence of input perturbations. In summary, the ablation study validates that each component of NQE-DR contributes indispensably to its overall robustness. Furthermore, the framework demonstrates appreciable generality, as it maintains competitive performance even when the core fully connected ansatz is replaced with a substantially different layered architecture.

Table 4. Performance comparison of NQE-DR and its ablation variants under different noise intensities and FGSM adversarial attacks. The top-performing result is highlighted in bold.

		Noise ϵ				FGSM δ			
Scheme	Clean	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
NQE-DR (Full)	**0.978**	**0.938**	**0.821**	**0.764**	**0.652**	**0.945**	**0.842**	**0.754**	**0.665**
NQE-DR-$\mathcal{T}$	0.945	0.890	0.753	0.680	0.610	0.921	0.805	0.650	0.602
NQE-DR-$\mathcal{U}_{\text{ent}}$	0.960	0.911	0.790	0.734	0.633	0.930	0.835	0.725	0.643
NQE-DR-$\mathcal{U}_{\text{layered}}$	0.968	0.925	0.805	0.745	0.640	0.938	0.838	0.740	0.655
NQE-DR-$\mathcal{R}$	0.963	0.895	0.761	0.710	0.625	0.919	0.814	0.680	0.614

5. Conclusion

This study proposes and examines a novel encoding strategy, NQE-DR, designed to enhance the robustness of quantum systems against adversarial attacks. This method addresses the common vulnerability of PQCs to minor adversarial perturbations. NQE-DR integrates nonlinear feature transformations during the encoding phase, alternating between R_Y and R_Z rotation gates, and a fully connected entanglement structure to reinforce expressiveness and resilience to subtle input variations while preserving the essential nonlinearity for complex learning tasks. A key innovation of NQE-DR is its dual regularisation strategy, which introduces dual constraints during optimisation. This mechanism enhances the circuit's resistance to attacks, leading to substantial gains in post-attack accuracy and providing a solid foundation for robustness. Comprehensive theoretical and empirical analyses demonstrate that NQE-DR consistently outperforms mainstream encoding approaches under clean, noisy, and adversarial conditions.

Although NQE-DR demonstrates superior robustness, further improvements are possible. Current experiments are limited to small qubit systems. Future research should explore larger-scale quantum systems and evaluate their computational stability. In addition to R_Y/R_Z, non-commuting gate sets (such as R_X or combined gates) can be evaluated to further enhance expressiveness. Embedding NQE-DR encoding into classical deep neural networks, such as convolutional neural networks and Transformers, could enable more complex hybrid quantum–classical architectures. Beyond classification tasks, NQE-DR can be applied to broader scenarios, including reinforcement learning, generative modelling, and quantum control.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/stars-art/quantum-robust [44].

Funding

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62 302 289 and 62 572 292) and the Science and Technology Innovation Plan of Shanghai Science and Technology Commission (Grant No. 23YF1416200).

ORCID iDs

YaoChong Li 0000-0002-1474-9800
XinXin Deng 0009-0002-6601-6278
RuiQing Xu 0000-0001-9064-6304
WenShan Xu 0000-0003-0088-0338
Ri-Gui Zhou 0000-0002-8894-8108

References

[1] Nielsen M A and Isaac L 2010 Chuang *Quantum Computation and Quantum Information: 10th Anniversary edn* (Cambridge University Press)
[2] Horowitz M and Grumbling E 2019 *Quantum Computing: Progress and Prospects* (National Academies Press)
[3] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N and Lloyd S 2017 Quantum machine learning *Nature* **549** 195–202
[4] Cerezo M, Verdon G, Huang H-Y, Cincio L and Coles P J 2022 Challenges and opportunities in quantum machine learning *Nat. Comput. Sci.* **2** 10
[5] Peral-García D, Cruz-Benito J and García-Peñalvo F J 2024 Systematic literature review: Quantum machine learning and its applications *Comput. Sci. Rev.* **51** 100619

[6] Li Y, Zhou R-G, Xu R, Luo J and Hu W 2020 A quantum deep convolutional neural network for image recognition *Quantum Sci. Technol.* **5** 044003
[7] Li Y, Qu Y, Zhou R-G and Zhang J 2025 Qmlsc: a quantum multimodal learning model for sentiment classification *Inf. Fusion* **120** 103049
[8] Wilkens S and Moorhouse J 2023 Quantum computing for financial risk measurement *Quantum Inf. Process.* **22** 51
[9] Herman D, Googin C, Liu X, Sun Y, Galda A, Safro I, Pistoia M and Alexeev Y 2023 Quantum computing for finance *Nat. Rev. Phys.* **5** 450–65
[10] Qu Z, Li Y and Tiwari P 2023 Qnmf: A quantum neural network based multimodal fusion system for intelligent diagnosis *Inf. Fusion* **100** 101913
[11] Wei L, Liu H, Xu J, Shi L, Shan Z, Zhao B and Gao Y 2023 Quantum machine learning in medical image analysis: a survey *Neurocomputing* **525** 42–53
[12] Gong W and Deng D-L 2021 universal adversarial examples and perturbations for quantum classifiers *Natl Sci. Rev.* **9** nwab130
[13] West M T, Tsang S-L, Low J S, Hill C D, Leckie C, Hollenberg L C L, Erfani S M and Usman M 2023 Towards quantum enhanced adversarial robustness in machine learning *Nat. Mach. Intell.* **5** 581–9
[14] West M T, Erfani S M, Leckie C, Sevior M, Hollenberg L C L and Usman M 2023 Benchmarking adversarially robust quantum machine learning at scale *Phys. Rev. Res.* **5** 023186
[15] Lu S, Duan L-M and Deng D-L 2020 Quantum adversarial machine learning *Phys. Rev. Res.* **2** 033212
[16] Ren W *et al* 2022 Experimental quantum adversarial learning with programmable superconducting qubits *Nat. Comput. Sci.* **2** 711–7
[17] Huang C and Zhang S 2023 Enhancing adversarial robustness of quantum neural networks by adding noise layers *New J. Phys.* **25** 083019
[18] Du Y, Hsieh M-H, Liu T, Tao D and Liu N 2021 Quantum noise protects quantum classifiers against adversaries *Phys. Rev. Res.* **3** 023153
[19] Gong W, Yuan D, Li W and Deng D-L 2024 Enhancing quantum adversarial robustness by randomized encodings *Phys. Rev. Res.* **6** 023020
[20] Maouaki W E, Marchisio A, Said T, Shafique M and Bennai M 2025 Designing robust quantum neural networks via optimized circuit metrics *Adv. Quantum Technol.* **8** 2400601
[21] Wendlinger M, Tscharke K and Debus P 2024 A comparative analysis of adversarial robustness for quantum and classical machine learning models *2024 IEEE Int. Conf. on Quantum Computing and Engineering (QCE)* vol 1 (IEEE) pp 1447–57
[22] Berberich J, Fink D, Pranjić D, Tutschku C and Holm C 2024 Training robust and generalizable quantum models *Phys. Rev. Res.* **6** 043326
[23] Wang C-C J and Bennink R S 2023 Variational quantum regression algorithm with encoded data structure (arXiv:2307.03334)
[24] Jiang H, Yang L, Bao Y, Rutong S, and Yang S 2024 Adaptive gradient regularization: a faster and generalizable optimization technique for deep neural networks (arXiv:2407.16944)
[25] Abbas A, Sutter D, Zoufal C, Lucchi A, Figalli A and Woerner S 2021 The power of quantum neural networks *Nat. Comput. Sci.* **1** 403–9
[26] Larocca M, Ju N, García-Martín D, Coles P J and Cerezo M 2023 Theory of overparametrization in quantum neural networks *Nat. Comput. Sci.* **3** 542–51
[27] Rath M and Date H 2024 Quantum data encoding: a comparative analysis of classical-to-quantum mapping techniques and their impact on machine learning accuracy *EPJ Quantum Technol.* **11** 72
[28] Kandala A, Mezzacapo A, Temme K, Takita M, Brink M, Chow J M and Gambetta J M 2017 Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets *Nature* **549** 242–6
[29] Nguyen T *et al* 2022 An evaluation of hardware-efficient quantum neural networks for image data classification *Electronics* **11** 437
[30] Ranga D, Prajapat S, Kumar P, Akhtar Z and Al-Khalidi M 2025 Quantum-enhanced classification: an empirical study of quantum support vector machines using the iris and wine datasets *Phys. Scr.* **100** 106006
[31] Pérez-Salinas A, Cervera-Lierta A, Gil-Fuster E and Latorre J I 2020 Data re-uploading for a universal quantum classifier *Quantum* **4** 226
[32] Goodfellow I J, Shlens J and Szegedy C 2015 Explaining and harnessing adversarial examples *3rd Int. Conf. on Learning Representations, (ICLR, San Diego, CA, USA, 7–9 May 2015, Conf. Track Proc.)*
[33] Cao K, Liu M, Su H, Wu J, Zhu J and Liu S 2021 Analyzing the noise robustness of deep neural networks *IEEE Trans. Vis. Comput. Graphics* **27** 3289–304
[34] Virmaux A and Scaman K 2018 Lipschitz regularity of deep neural networks: analysis and efficient estimation *Advances in Neural Information Processing Systems* vol 31 (Curran Associates, Inc)
[35] Fazlyab M, Robey A, Hassani H, Morari M and Pappas G J 2019 Efficient and accurate estimation of lipschitz constants for deep neural networks *Proc. 33rd Int. Conf. on Neural Information Processing Systems, (Red Hook, NY, USA, Curran Associates Inc)*
[36] Pauli P, Koch A, Berberich J, Kohler P and Allgöwer F 2021 Training robust neural networks using lipschitz bounds *IEEE Control Syst. Lett.* **6** 121–6
[37] Yongxi Y, Shibin Z, Lili Y and Yan C 2024 Hybrid classical quantum neural network with high adversarial robustness *Proc. 2024 Int. Conf. on Machine Learning and Intelligent Computing (Proc. Machine Learning Research)* vol 245 (PMLR) pp 271–9
[38] Sim S, Johnson P D and Aspuru-Guzik A 2019 Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms *Adv. Quantum Technol.* **2** 1900070
[39] Ballarin M, Mangini S, Montangero S, Macchiavello C and Mengoni R 2023 Entanglement entropy production in quantum neural networks *Quantum* **7** 1023
[40] Kingma D P and Jimmy B 2015 Adam: a method for stochastic optimization *3rd Int. Conf. on Learning Representations (Conf. Track Proc., ICLR 2015, San Diego, CA, USA, 7–9 May 2015)*
[41] Bergholm V 2018 et al Pennylane: automatic differentiation of hybrid quantum-classical computations (arXiv:1811.04968)
[42] Rocklin M 2015 Dask: parallel computation with blocked algorithms and task scheduling *Proc. 14th Python in Science Conf.* pp 130–6
[43] Cerezo M *et al* 2021 Variational quantum algorithms *Nat. Rev. Phys.* **3** 625–44
[44] Y Li *et al* 2025 Dual-regularized nonlinear quantum encoding for adversarial robustness in quantum machine learning (available at: https://github.com/stars-art/quantum-robust)
[45] Nakaji K and Yamamoto N 2021 Expressibility of the alternating layered ansatz for quantum computation *Quantum* **5** 434