

CONTENTS

I. Theoretical Background	1
A. Supervised learning	1
B. Variational quantum learning models	2
II. The Automated Learning Strategy	3
A. Choose an appropriate number of qubits	3
B. Encode data into unitaries	4
C. Training process and efficient compiling of U_y	6
D. Prediction and evaluation	6
E. Gradient perspective	6
III. Physical Interpretation and Analysis of Training Process	7
A. Formulation of the training process	7
B. Convergence to global minimum	9
C. Convergence with constant probability	9
D. Heavy-tailed Hamiltonian	11
IV. Other Properties of Quantum Automated Learning	11
A. Generalization	11
B. Universal representation power	12
C. State reusability	13
D. Mini-batch optimization	14
References	14

I. THEORETICAL BACKGROUND

A. Supervised learning

We start by introducing the basic framework of supervised learning [1]. Let \mathcal{X} be the set of input data and $\mathcal{Y} = \{1, 2, \dots, k\}$ be the set of labels. We assume that every input data $\mathbf{x} \in \mathcal{X}$ has a deterministic label $y(\mathbf{x}) \in \mathcal{Y}$. Let \mathcal{D} be an unknown distribution over \mathcal{X} . The goal of supervised learning is to find an algorithm $\mathcal{A}(\cdot)$ (probably randomized in quantum machine learning) such that, input a sample $\mathbf{x} \sim \mathcal{D}$, output the label $y(\mathbf{x})$ with high probability. To achieve this goal, we parametrize the learning model by parameters θ and optimize the average loss

$$R(\theta) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} L(\mathbf{x}, y(\mathbf{x}); \theta). \quad (\text{S1})$$

Here $L(\mathbf{x}, y; \theta)$ is some loss function, usually a metric of the difference between the output distribution of $\mathcal{A}(\mathbf{x}; \theta)$ and the correct label y . $R(\theta)$ is called the risk or the prediction error of the model $\mathcal{A}(\cdot; \theta)$. However, the distribution \mathcal{D} is unknown, so we cannot directly calculate $R(\theta)$. Instead, we sample a training dataset $S = \{(\mathbf{x}_i, y_i = f(\mathbf{x}_i))\}_{i=1}^m$ from \mathcal{D} , and optimize the following empirical risk or training error:

$$\hat{R}_S(\theta) = \frac{1}{m} \sum_{i=1}^m L(\mathbf{x}_i, y_i; \theta). \quad (\text{S2})$$

According to the simple decomposition $R(\theta) = \hat{R}_S(\theta) + (R(\theta) - \hat{R}_S(\theta))$, the success of supervised learning depends on two important factors: trainability and generalization. In short, trainability asks whether we can efficiently find θ with low empirical risk, while generalization asks whether the generalization gap $\text{gen}_S(\theta) = R(\theta) - \hat{R}_S(\theta)$ is upper bounded, i.e., whether the good performance on the training set S can be generalized to unseen data.

B. Variational quantum learning models

For conventional gradient-based quantum learning approaches [2], a learning algorithm $\mathcal{A}(\mathbf{x}; \boldsymbol{\theta})$ executes a variational quantum circuit $V(\boldsymbol{\theta})$ to a data-encoded state $|\phi(\mathbf{x})\rangle = U(\mathbf{x})|0^n\rangle$ before performing certain measurements to make the prediction. Assuming the measurement observable to be O_M , the output from the variational circuit is the expectation value $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \phi(\mathbf{x}) | V(\boldsymbol{\theta})^\dagger O_M V(\boldsymbol{\theta}) | \phi(\mathbf{x}) \rangle = \langle 0^n | U(\mathbf{x})^\dagger V(\boldsymbol{\theta})^\dagger O_M V(\boldsymbol{\theta}) U(\mathbf{x}) | 0^n \rangle$. The loss function is often defined as a function of this value, where commonly used forms include mean square error and cross-entropy. For a training task, the average loss value over a given set of training data is defined as the empirical risk, where schemes based on gradient descents are widely exploited to minimize it and find the optimal parameters $\boldsymbol{\theta}^*$. In the quantum machine learning realm, there are various methods proposed to calculate the gradients with respect to circuit parameters, including finite differences, the parameter-shift rules, and quantum natural gradients [3–5].

Quantum neural networks have demonstrated promising generalization capabilities in various learning settings [6, 7]. Intuitively, when the number of training data points exceeds the degrees of freedom in the parameter space, the generalization gap of the optimized parameters is typically bounded by a small constant. However, the practical trainability of quantum neural networks faces two significant challenges: the prohibitive cost of estimating gradients and the pathological landscape of loss function.

Computation cost of gradient estimation. A key bottleneck lies in the computational cost of estimating gradients with respect to the circuit parameters. In gradient-based optimization, the parameter-shift rule is among the most commonly used techniques for computing quantum gradients [3, 4]. For a circuit parameter θ_i associated with a gate of the form $e^{-i\theta_i P/2}$, where P is a Pauli operator, the exact gradient can be expressed as

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_i} = \frac{1}{2} \left[f(\mathbf{x}, \boldsymbol{\theta} + \frac{\pi}{2} \mathbf{e}_i) - f(\mathbf{x}, \boldsymbol{\theta} - \frac{\pi}{2} \mathbf{e}_i) \right],$$

where \mathbf{e}_i is the unit vector whose i -th entry is 1. Thus, evaluating a single gradient component requires two circuit executions with shifted parameter values. However, each expectation value $f(\mathbf{x}, \boldsymbol{\theta} \pm \frac{\pi}{2} \mathbf{e}_i)$ must itself be estimated statistically from repeated circuit measurements. Obtaining an accurate estimate of f typically requires on the order of thousands of circuit repetitions even for a single data sample. As the number of parameters increases, this cost quickly becomes prohibitive. To make this comparison more explicit, let T denote the number of training steps, n_p the number of trainable parameters, and M the number of circuit repetitions required to estimate one expectation value. The total number of network or circuit runs for different learning paradigms can be summarized as follows:

- **Classical neural networks:** Owing to the backpropagation algorithm, each training step requires roughly one forward and one backward pass through the network (the backward pass is more expensive but only by a constant factor). Hence, the total number of network runs scales as $O(T)$.
- **Quantum neural networks:** As discussed above, each training step requires estimating the gradient with respect to every parameter, and each gradient evaluation involves $2M$ circuit executions. The total number of circuit runs therefore scales as $2MTn_p$.
- **Quantum automated learning (QAL):** As will be introduced below, our QAL protocol requires running the training circuit only $1/p$ times on average, where p is the success probability of post-selection. Although the depth of the training circuit is proportional to T , the overall computational cost scales as $O(T/p)$, without an explicit dependence on n_p or any additional multiplicative factor M .

Even for moderate problem sizes (e.g., $T \sim 10^4$, $n_p \sim 10^5$, and $M \sim 10^3$), the total number of circuit executions in a quantum neural network already reaches the order of 10^{12} , let alone for larger-scale models with more parameters or longer training schedules. This reveals a sobering fact that, even if we now have a large fault-tolerant quantum computer, even if there is no barrier like barren plateaus in the training step, we still cannot deploy conventional quantum neural networks for large models!

In a recent paper [8], the authors advocate prioritizing scalable training as a central objective in quantum machine learning. We find this viewpoint closely aligned with our own motivation, so we briefly outline it here. The success of deep learning have emphasized the importance of scaling. However, the vast majority of variational quantum machine learning studies focus mainly on understanding and encoding inductive biases in quantum models, while the issue of scalability is often overlooked or reduced to the discussion of barren plateaus. In particular, little attention has been given to the probably more fundamental barrier to scalability: the inherently high cost of extracting gradient information from quantum circuits [9]. Since scalable models seems to be both necessary and rare, it might be better to build scalable models first, and only then focus on how to improve the performance by encoding problem-specific biases into the models.

Our QAL protocol follows this idea. The training cost of our model is roughly T/p , where T is the number of training step and p is the post-selection success probability. This does not explicitly suffer from the Mn_p scaling, thus it is possible to deploy it for large models. Of course, the overall advantage depends on how the post-selection success probability p scales with the problem size, for which we give a justification in Sec. III.

Pathological landscape of loss functions Another barrier to scalability lies in the landscape of loss functions. The loss landscape of quantum neural networks can be highly non-convex and difficult to navigate. As shown in Ref. [10], the loss function of quantum neural networks exhibits exponentially many local minima, which can easily trap optimization algorithms and hinder convergence. In parallel, the phenomenon of barren plateaus, first identified in Ref. [11], poses a critical issue: the gradients of the loss function tend to vanish exponentially with the number of qubits, especially in deep quantum circuits. In such cases, the loss landscape becomes effectively flat, making it extremely difficult to identify a meaningful direction for optimization. The presence of barren plateaus is closely related to the randomness and entanglement structure of the quantum circuit, as well as the choice of cost function and parameter initialization, all of which severely threaten the scalability and practical trainability of quantum neural networks for large-scale problems [11–16].

Our QAL framework circumvents this difficulty by stepping out of the parameter space and optimizing directly in the Hilbert space. To better illustrate this distinction, it is instructive to compare QAL with the so-called *flipped variational quantum circuits*.

In conventional approaches, the data-encoding circuit $U(x)$ is applied before the variational circuit $V(\theta)$, which mirrors the standard machine learning framework where input data are fed into the model. However, there is no a priori reason that $U(x)$ must precede $V(\theta)$. Recent studies have explored alternative structures, such as data re-uploading [17] (where data encoding and variational layers are interleaved) and data-parameter combining [18] (where the rotation angles in the circuit are functions of both data and parameters, e.g., $x + \theta$). In particular, Ref. [19] introduced the *flipped model*, where the order of $V(\theta)$ and $U(x)$ is swapped. They demonstrated that such architecture can still exhibit quantum advantage under the computational assumption that the discrete logarithm is hard for classical computers. Therefore, we expect the flipped model to have comparable learning capabilities to the conventional one. More importantly, the flipped model shares a closer conceptual connection with our QAL framework, which we elaborate on below.

Definition S1 (Flipped model, adapted from [19]). *Let $V(\theta)$ be a variational quantum circuit, $U(x)$ an encoding quantum circuit, and $O(x)$ a data-dependent observable. A flipped model is defined by the parametrized function:*

$$f(x; \theta) = \langle 0^n | V(\theta)^\dagger U(x)^\dagger O(x) U(x) V(\theta) | 0^n \rangle. \quad (\text{S3})$$

Suppose the goal is to minimize $f(x; \theta)$. For example, in a classification task, one may take $O(x) = \mathbf{I} - \Pi_{y(x)}$, the projection operator onto the incorrect outputs, so that $f(x; \theta)$ represents the classification error. The training loss is then the empirical expectation $\mathbb{E}_{x \sim S}[f(x; \theta)]$ (S is the training dataset and $x \sim S$ means that x is uniformly drawn from S), which can be written as

$$\mathbb{E}_{x \sim S}[f(x; \theta)] = \langle \theta | H_S | \theta \rangle, \quad \text{where } |\theta\rangle = V(\theta) | 0^n \rangle, \quad H_S = \mathbb{E}_{x \sim S}[U(x)^\dagger O(x) U(x)].$$

Remarkably, this loss takes exactly the same quadratic form as in QAL. Hence, the training objective of the flipped model is also equivalent to finding the ground state of H_S . The key difference lies in the optimization approach: in the variational quantum circuit framework, one searches for the ground state by adjusting parameters within the manifold $\{|\theta\rangle = V(\theta) | 0^n \rangle\}_\theta$ using gradient-based optimization. This approach is therefore constrained by the chosen ansatz $V(\theta)$ and susceptible to the pathological loss landscape issues discussed above. In contrast, QAL bypasses this limitation by jumping out of the parameter manifold and performing optimization directly in the Hilbert space through a physical process, such as imaginary-time evolution, to prepare the ground state. This direct optimization strategy avoids the pathological loss landscapes and provides a scalable path toward training quantum models.

II. THE AUTOMATED LEARNING STRATEGY

In this section, we provide more technical details about the quantum automated learning strategy.

A. Choose an appropriate number of qubits

To carry out the QAL protocol, the first step is to decide an appropriate number of qubits n . Since an n -qubit state lives in a 2^n -dimensional Hilbert space and thus bears $\Theta(2^n)$ degrees of freedom, one natural choice is $n = \Theta(\log |x|)$, where $|x|$ is the dimension of data sample x . However, we remark that the choice of n is much more flexible. For example, if we are classifying Hamiltonian data, it is more natural to set n to be the system size of the Hamiltonian. If the data are images of size $L \times L$, setting $n = L$ may align with the two-dimensional structure better.

B. Encode data into unitaries

Once we pin down the number of qubits n , the next step is to encode data into n -qubit unitaries. Here we present the detailed data encoding schemes for quantum automated learning, which incorporates three distinct categories of data: classical data, Hamiltonian data, and quantum state data. An overview of the encoding methods is provided in Fig. S1, which summarizes the key approaches before delving into the detailed descriptions of each scheme. We note that other schemes can also be explored.

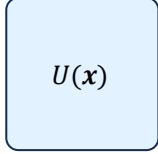
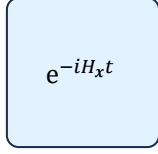
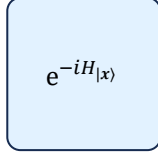
	Classical data	Hamiltonian data	Quantum state data
Original data	x	H_x	$ x\rangle$
Encoding unitary	Parameterized quantum circuit  $U(x)$	Real-time evolution  $e^{-iH_x t}$	$H_{ x\rangle} = (\langle x \otimes \mathbf{I}_n)H(x\rangle \otimes \mathbf{I}_n)$  $e^{-iH_{ x\rangle}}$

FIG. S1: **Overview of three different data encoding methods.** We encode classical data x into parameterized quantum circuit $U(x)$; encode Hamiltonian data H_x into its real-time evolution $e^{-iH_x t}$; and encode quantum state $|x\rangle$ first into an n -qubit Hamiltonian: $H_{|x\rangle} = (\langle x| \otimes \mathbf{I}_n)H(|x\rangle \otimes \mathbf{I}_n)$, and then encode it to a unitary $e^{-iH_{|x\rangle}}$.

Classical data, including images, text, or audio, can be transformed into a vector of numerical values, denoted as x . The data vector x is encoded into parameterized quantum circuit. Specifically, each element of x is mapped into rotation angles of single-qubit gates. A single-qubit gate is parametrized as $G(\alpha, \beta, \gamma) = R_y(\alpha)R_z(\beta)R_y(\gamma)$, where $R_y(\alpha)$ and $R_z(\beta)$ are the rotations around the Y and Z axes of the Bloch sphere by angle α and β , respectively. Therefore, for a n -qubit quantum circuit, a layer of single-qubit gates can encode up to $3n$ entries of the vector x . If we denote the dimension of x as l , then it is necessary to employ $\lceil \frac{l}{3n} \rceil$ layers of single-qubit gates. More concretely, considering a $3n$ -dimensional vector y , we define the encoding of a single-qubit layer as: $G(y) = \bigotimes_{i=1}^n G_i(y_{2n+i}, y_{n+i}, y_i)$, where G_i acts on the i -th qubit, as illustrated in Fig. S2a. Then the k -th layer single-qubit encoding of the data vector x is defined as $G(x_{3n(k-1)+1:3nk})$, where $x_{i:j}$ denotes the abbreviation of $(x_i, x_{i+1}, \dots, x_j)$. In cases where the number of elements in x does not exactly divide by $3n$, padding with zeros is used to ensure uniformity.

Between two layers of single-qubit gates, we insert a layer of two-qubit gates to entangle the qubits, leading to the spread of information. This layer of two-qubit gates is composed of a CNOT-gate block A and a CZ-gate block B . Each block consists of two layers of two-qubit gates: in the first layer, the odd-numbered qubits act as the control qubits, while in the second layer, the even-numbered qubits serve as the control qubits. In both layers, each control qubit targets the subsequent qubit in the sequence. Mathematically, we define: $A = \left(\bigotimes_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} \text{CNOT}_{2i, 2i+1} \right) \left(\bigotimes_{i=1}^{\lfloor \frac{n}{2} \rfloor} \text{CNOT}_{2i-1, 2i} \right)$ and $B = \left(\bigotimes_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} \text{CZ}_{2i, 2i+1} \right) \left(\bigotimes_{i=1}^{\lfloor \frac{n}{2} \rfloor} \text{CZ}_{2i-1, 2i} \right)$, as shown in Fig. S2a. To ensure that all elements of the data vector can influence the measured qubits used for prediction, we add additional $\lfloor \frac{n}{2} \rfloor - 1$ layers of two-qubit gates.

The final unitary encoding $U(x)$ for classical data is then given by a sequence of single- and two-qubit gates:

$$(BA)^{\lfloor \frac{n}{2} \rfloor - 1} G(x_{3n(d-1)+1:3nd}) \cdots BAG(x_{3n+1:6n}) BAG(x_{1:3n}), \quad (\text{S4})$$

where $d = \lceil \frac{l}{3n} \rceil$ and x is padded with zeros if $3nd > l$, as illustrated in Fig. S2b.

For Hamiltonian data, we encode the Hamiltonian H_x through its real-time evolution $e^{-iH_x t}$. This encoding inherently captures the time evolution of quantum states governed by the Schrödinger equation. Both $e^{-iH_x t}$ and its reverse evolution $e^{iH_x t}$ can be implemented through quantum Hamiltonian simulation techniques [20–23]. One may exploit other encoding schemes for H_x . In practice, we find that our real-time evolution encoding works well, as shown in our numerical simulations.

For quantum state classification, each datum is a quantum state $|x\rangle$ of s qubits. In order to carry out the QAL protocol for classifying $|x\rangle$, we need to encode $|x\rangle$ into a unitary $U_{|x\rangle}$. We fix an $(s+n)$ -qubit Hamiltonian H . We first encode the quantum state $|x\rangle$ into an n -qubit Hamiltonian: $H_{|x\rangle} = (\langle x| \otimes \mathbf{I}_n)H(|x\rangle \otimes \mathbf{I}_n)$, and then encode it to a unitary $U_{|x\rangle} = e^{-iH_{|x\rangle}}$. This unitary can be efficiently implemented using copies of $|x\rangle$ and real-time evolution $e^{-iH t}$, inspired by the Lloyd-Mohseni-Rebentrost protocol [24] that implements $e^{-i\rho t}$ from copies of ρ . More concretely, for any state ρ straightforward calculations

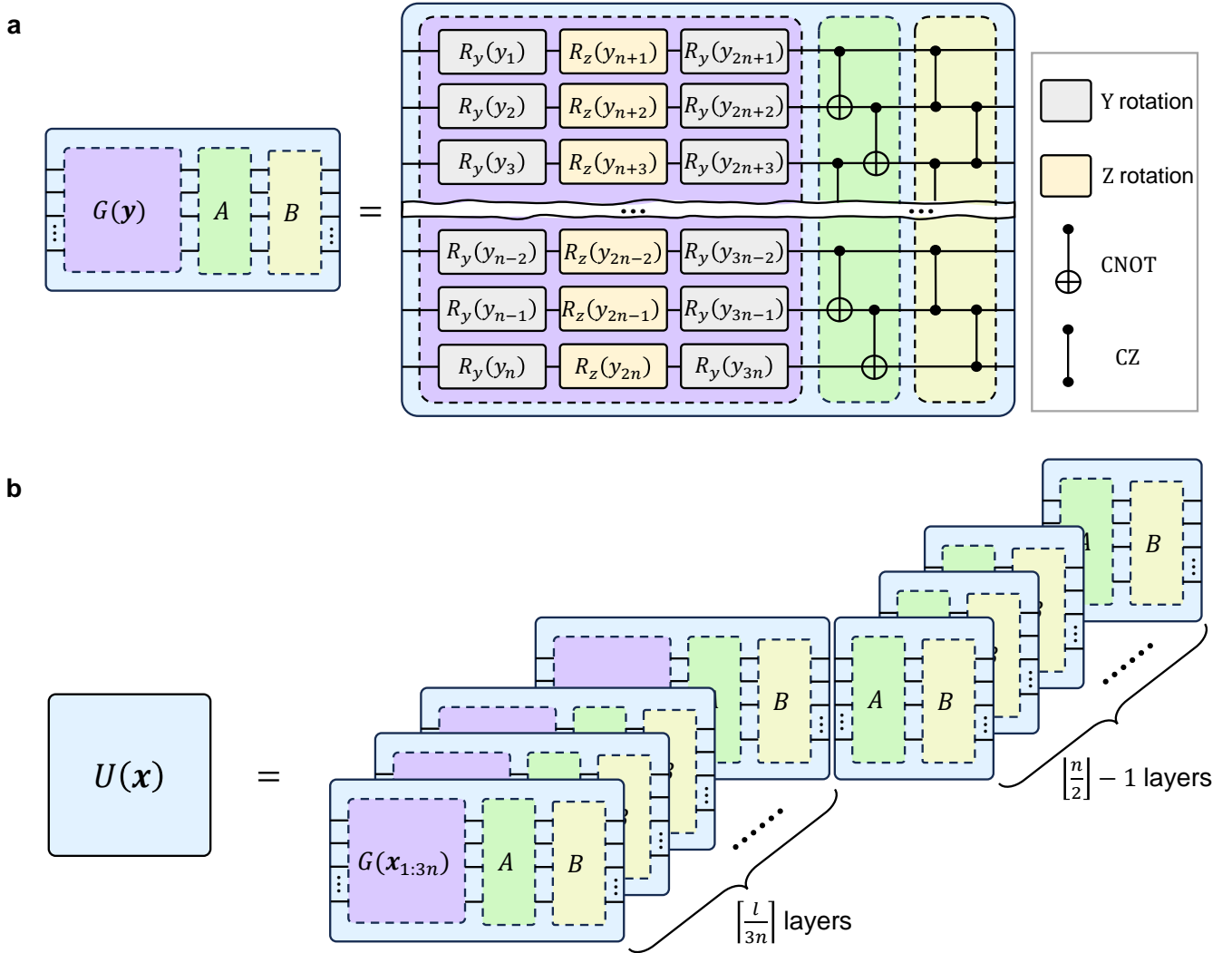


FIG. S2: **The classical data encoding scheme.** **a**, Illustrates an encoding layer for a $3n$ -dimensional vector \mathbf{y} ($G(\mathbf{y})$). The layer consists of three layers of single-qubit gates (R_y, R_z, R_y), denoted by the purple block; two layers of CNOT two-qubit gates, denoted by the green block; and two layers of CZ two-qubit gates, denoted by the yellow block. **b**, Illustrates the complete encoding for an l -dimensional vector \mathbf{x} . The encoding scheme involves $\lceil \frac{l}{3n} \rceil$ layers of single-qubit and two-qubit gates ($BAG(\mathbf{x}_{3n(k-1)+1:3nk})$), followed by $\lfloor \frac{n}{2} \rfloor - 1$ layers of entangling gates (BA).

166 yield

$$\begin{aligned}
 \text{tr}_{\leq l}(e^{-iH\Delta t}(|\mathbf{x}\rangle\langle\mathbf{x}| \otimes \rho)e^{iH\Delta t}) &= \rho + \text{tr}_{\leq l}(-iH\Delta t(|\mathbf{x}\rangle\langle\mathbf{x}| \otimes \rho) + (|\mathbf{x}\rangle\langle\mathbf{x}| \otimes \rho)iH\Delta t) + O((\Delta t)^2) \\
 &= \rho - i\Delta t[H_{|\mathbf{x}\rangle}, \rho] + O((\Delta t)^2) \\
 &= e^{-iH_{|\mathbf{x}\rangle}\Delta t}\rho e^{iH_{|\mathbf{x}\rangle}\Delta t} + O((\Delta t)^2).
 \end{aligned} \tag{S5}$$

167 Therefore, if we apply $e^{-iH\Delta t}$ to $|\mathbf{x}\rangle\langle\mathbf{x}| \otimes \rho$, we effectively apply $e^{-iH_{|\mathbf{x}\rangle}\Delta t}$ to ρ , up to a second-order error $O((\Delta t)^2)$.
 168 Repeating this procedure $1/\Delta t$ times, we effectively apply $e^{-iH_{|\mathbf{x}\rangle}}$ to ρ up to error $O(\Delta t)$. By choosing Δt sufficiently small,
 169 we can approximate $e^{-iH_{|\mathbf{x}\rangle}}$ up to any precision. In our numerical simulations, we choose H as a random 4-local Hamiltonian.
 170 Each term of H is a 4-body Pauli interaction on 4 random positions with a random interaction strength between -1 and 1.

C. Training process and efficient compiling of U_y

In this subsection, we show that the required unitary U_y to implement the target-oriented perturbation can be compiled in an efficient way, with the count of CNOT gates scales logarithmically with the number of classes k . Recalling that $U_y = M_y \otimes Z + \sqrt{I - M_y^2} \otimes X$ acts on $\lceil \log k \rceil + 1$ qubits, and $M_y = |y\rangle\langle y| + (1 - \eta)(\mathbf{I} - |y\rangle\langle y|)$, we arrive at the following decomposition:

$$\begin{aligned} U_y &= |y\rangle\langle y| \otimes Z + (I - |y\rangle\langle y|) \otimes \left((1 - \eta)Z + \sqrt{2\eta - \eta^2}X \right) \\ &= \left[I \otimes \left((1 - \eta)Z + \sqrt{2\eta - \eta^2}X \right) \right] \left[|y\rangle\langle y| \otimes \left((1 - \eta)Z + \sqrt{2\eta - \eta^2}X \right) Z + (I - |y\rangle\langle y|) \otimes I \right]. \end{aligned} \quad (\text{S6})$$

The first part of the above equation $I \otimes \left((1 - \eta)Z + \sqrt{2\eta - \eta^2}X \right)$ is a single-qubit gate and can be compiled into a constant number of gates. The second part is a $\lceil \log k \rceil$ -controlled $SU(2)$ gate, as $|y\rangle\langle y|$ involves $\lceil \log k \rceil$ qubits and $\left((1 - \eta)Z + \sqrt{2\eta - \eta^2}X \right) Z$ is a $SU(2)$ gate. Utilizing the techniques in Ref. [25], such a multi-controlled $SU(2)$ gate can be compiled into $O(\log k)$ CNOT gates. This leads to the conclusion that the whole U_y can be compiled in an efficient way with about $\log k$ two-qubit gates.

D. Prediction and evaluation

The predicted label of \mathbf{x} is the outcome of measurements in the computational basis performed on $U(\mathbf{x})|\psi\rangle$. So the probability of correct prediction (i.e., the accuracy) is $\langle \psi | U(\mathbf{x})^\dagger \Pi_{y(\mathbf{x})} U(\mathbf{x}) | \psi \rangle = 1 - \langle \psi | H_{\mathbf{x}} | \psi \rangle$. This accuracy can be amplified by repetition. Indeed, once the data in each step has been sampled, the training process is a fixed quantum circuit (with post-selection). So we can run the circuit multiple times to obtain K copies of the final states $|\psi\rangle$. To predict the label of a new unseen data sample \mathbf{x} , we measure $U(\mathbf{x})|\psi\rangle$ in the computational basis for each copy and do a majority vote. For simplicity, we consider the binary classification problem and assume K is odd, then the probability of correct label (called K -accuracy) is

$$p_K(\mathbf{x}, |\psi\rangle) = \sum_{r=0}^{(K-1)/2} \binom{K}{r} \langle \psi | H_{\mathbf{x}} | \psi \rangle^r (1 - \langle \psi | H_{\mathbf{x}} | \psi \rangle)^{K-r}.$$

When $K = 1$, this reduces to the single-copy accuracy $1 - \langle \psi | H_{\mathbf{x}} | \psi \rangle$. When $K = \infty$, the K -accuracy equals to the step function $\mathbf{1}[\langle \psi | H_{\mathbf{x}} | \psi \rangle < 1/2]$. Throughout this paper, we call K the number of trials.

E. Gradient perspective

As mentioned in the main text, the dissipation process of quantum automated learning actually implements the gradient descent algorithm in an automated manner. In the training step (iii), the quantum system is evolved through $U(\mathbf{x})$, a y -dependent perturbation M_y and $U(\mathbf{x})^\dagger$. The perturbation is described by the local operator $M_y = |y\rangle\langle y| + (1 - \eta)(\mathbf{I} - |y\rangle\langle y|)$, which, as detailed in the Methods, is confined to the measurement subsystem used for the classification label. To formalize this on the entire n -qubit system, we define the global operator $\mathcal{M}_y = \Pi_{y(\mathbf{x})} + (1 - \eta)(\mathbf{I} - \Pi_{y(\mathbf{x})})$. This operator is constructed to implement the local perturbation M_y on the label subsystem while acting as the identity on all other complementary qubits. Here, $\Pi_{y(\mathbf{x})}$ is the full-space measurement projection corresponding to the state encoding the label $y(\mathbf{x})$, which allows us to define the Hamiltonian $H_{\mathbf{x}} = \mathbf{I} - U(\mathbf{x})^\dagger \Pi_{y(\mathbf{x})} U(\mathbf{x})$. The resulting (unnormalized) state after this step is $U(\mathbf{x})^\dagger \mathcal{M}_y U(\mathbf{x})|\psi\rangle = (\mathbf{I} - \eta H_{\mathbf{x}})|\psi\rangle$. As mentioned in Methods, the non-unitary perturbation M_y in training step (iii) of the QAL protocol is implemented by block encoding into a unitary U_y with an ancillary qubit combined with post-selection. As a result, this step effectively updates the state with unitary transformation:

$$|\psi\rangle \leftarrow \frac{(\mathbf{I} - \eta H_{\mathbf{x}})|\psi\rangle}{\|(\mathbf{I} - \eta H_{\mathbf{x}})|\psi\rangle\|}, \quad (\text{S7})$$

where $\|(\mathbf{I} - \eta H_{\mathbf{x}})|\psi\rangle\|$ is a normalization factor whose square gives the success probability of post-selection.

The probability of correct prediction of a datum \mathbf{x} reads $\langle \psi | U(\mathbf{x})^\dagger \Pi_{y(\mathbf{x})} U(\mathbf{x}) | \psi \rangle = 1 - \langle \psi | H_{\mathbf{x}} | \psi \rangle$. As mentioned in the main text, we define the loss function as the average failure probability: $\hat{R}_S(\psi) = \mathbb{E}_{\mathbf{x} \sim S} \langle \psi | H_{\mathbf{x}} | \psi \rangle$, where $\mathbb{E}_{\mathbf{x} \sim S}$ denotes the expectation and $\mathbf{x} \sim S$ means \mathbf{x} is uniformly sampled from the training set S . From the perspective of conventional machine learning, we may also regard $|\psi\rangle$ as a variational state parametrized by a complex vector ψ . Given that the expectation value

207 $\langle \psi | H_{\mathbf{x}} | \psi \rangle$ is real, we transform the complex vector $\psi = (\psi_1, \dots, \psi_{2^n})$ into a fully real representation: $(a_1, b_1, \dots, a_{2^n}, b_{2^n})$,
 208 where a_i and b_i denote the real and imaginary components of ψ_i respectively. Owing to the Hermitian property of $H_{\mathbf{x}}$, we
 209 derive the partial derivatives: $\frac{\partial \langle \psi | H_{\mathbf{x}} | \psi \rangle}{\partial a_i} = 2 \operatorname{Re} \left(\sum_j (H_{\mathbf{x}})_{i,j} \psi_j \right)$ and $\frac{\partial \langle \psi | H_{\mathbf{x}} | \psi \rangle}{\partial b_i} = 2 \operatorname{Im} \left(\sum_j (H_{\mathbf{x}})_{i,j} \psi_j \right)$. This allows us to
 210 define $\frac{\partial \langle \psi | H_{\mathbf{x}} | \psi \rangle}{\partial \psi_i} = 2 \sum_j (H_{\mathbf{x}})_{i,j} \psi_j$. Consequently, the gradient of $\langle \psi | H_{\mathbf{x}} | \psi \rangle$ with respect to ψ can be succinctly expressed
 211 as $2H_{\mathbf{x}} |\psi\rangle$. Therefore, the update rule in Eq. (S7) essentially implements the stochastic projected gradient descent algorithm
 212 to minimize the loss function $\hat{R}_S(\psi)$ with a batch size one. Here we use the term “projected” to emphasize the normalization
 213 after each update. From the stochastic gradient descent perspective, one may conclude that an initial state $|\psi\rangle$ can exponentially
 214 converge to a local minimum through updating rule (S7) on expectation [26]. However, a rigorous proof of convergence to
 215 the global minimum is unattainable in general. In fact, this is an inherent drawback for conventional gradient-based quantum
 216 learning approaches. Whereas, owing to the quadratic form of the loss function and the clear physical interpretation, we can
 217 rigorously prove that $|\psi\rangle$ converges exponentially to the global minimum for the QAL protocol, as discussed in the main text
 218 and detailed in the following sections.

219 III. PHYSICAL INTERPRETATION AND ANALYSIS OF TRAINING PROCESS

220 Throughout this section, we use $\|A\|_1, \|A\|_\infty$ to denote the trace norm (the summation of singular values) and spectral norm
 221 (the largest singular value), respectively. For two Hermitian A, B , denote $A \preceq B$ if $B - A$ is positive semi-definite. By definition,
 222 for any \mathbf{x} , $0 \preceq H_{\mathbf{x}} \preceq I$, and thus $0 \preceq H_S \preceq I$. We will use the following fact.

223 **Lemma S1.** *Let A, B, C be three Hermitian matrices such that $\|A\|_1 \leq 1$, $0 \preceq B, C \preceq I$. Then*

$$\|BAB\|_1 \leq 1, \|BAC + CAB\|_1 \leq 2. \quad (\text{S8})$$

224 *Proof.* We first prove the lemma when A is a normalized pure state $|a\rangle\langle a|$. Write $|b\rangle = B|a\rangle, |c\rangle = C|a\rangle$. Since $0 \preceq B, C \preceq I$,
 225 the norms of $|b\rangle, |c\rangle$ are at most 1. Then $\|BAB\|_1 = \||b\rangle\langle b|\|_1 \leq 1$, $\|BAC + CAB\|_1 = \||b\rangle\langle c| + |c\rangle\langle b|\|_1 \leq 2$. For general
 226 A , write the spectrum decomposition $A = \sum_i \lambda_i |\lambda_i\rangle\langle \lambda_i|$. By triangle inequality, $\|BAB\|_1 \leq \sum_i |\lambda_i| \|B|\lambda_i\rangle\langle \lambda_i|B\|_1 \leq$
 227 $\sum_i |\lambda_i| = \|A\|_1 \leq 1$ and $\|BAC + CAB\|_1 \leq \sum_i |\lambda_i| \|B|\lambda_i\rangle\langle \lambda_i|C + C|\lambda_i\rangle\langle \lambda_i|B\|_1 \leq \sum_i 2|\lambda_i| \leq 2$. \square

228 A. Formulation of the training process

229 In this subsection, we explain the training process from a physical perspective and derive an analytical characterization of the
 230 success probability of post-selection and the performance of the final model. Observe that the empirical risk is the energy of $|\psi\rangle$
 231 under the Hamiltonian H_S , so finding the global minimum is equivalent to finding the ground state of H_S . We rewrite (S7) in
 232 the density matrix formalism:

$$\rho \leftarrow (I - \eta H_{\mathbf{x}}) \rho (I - \eta H_{\mathbf{x}}). \quad (\text{S9})$$

233 Here we keep the post-state ρ unnormalized. Indeed, $\operatorname{Tr}(\rho)$ is the success probability of the post-selection. So the density matrix
 234 formalism helps us to keep track of the overall success probability. Another benefit of the density matrix formalism is that we
 235 can embed the randomness of the sample into the state. Since the datum \mathbf{x} is uniformly sampled from S , the averaged post-state
 236 up to the second order term is

$$\begin{aligned} \rho &\leftarrow \mathbb{E}_{\mathbf{x} \sim S} (I - \eta H_{\mathbf{x}}) \rho (I - \eta H_{\mathbf{x}}) \\ &\approx \rho - \eta (H_S \rho + \rho H_S) \\ &\approx e^{-\eta H_S} \rho e^{-\eta H_S}. \end{aligned} \quad (\text{S10})$$

237 We make this approximation precise in the following lemma

Lemma S2.

$$\mathbb{E}_{\mathbf{x} \sim S} (I - \eta H_{\mathbf{x}}) \rho (I - \eta H_{\mathbf{x}}) = e^{-\eta H_S} \rho e^{-\eta H_S} + \eta^2 O, \quad (\text{S11})$$

238 where O is a Hermitian matrix with trace norm at most 4.

239 *Proof.* Let $R = (e^{-\eta H_S} - (I - \eta H_S))/\eta^2$. Since $0 \preceq H_S \preceq I$, all the eigenvalues of H_S are in $[0, 1]$. By the Taylor expansion
 240 of the exponential function, for any $x \in [0, 1]$, there exists $x^* \in [0, 1]$ such that $(e^{-\eta x} - (1 - \eta x))/\eta^2 = (x^*)^2/2 \in [0, 1/2]$.
 241 Therefore, R is a Hermitian matrix such that $0 \preceq R \preceq I/2$. By Lemma S1, we have

$$\begin{aligned} & \|\mathbb{E}_{\mathbf{x} \sim S}(I - \eta H_{\mathbf{x}})\rho(I - \eta H_{\mathbf{x}}) - e^{-\eta H_S}\rho e^{-\eta H_S}\|_1 \\ &= \|\rho - \eta(H_S\rho + \rho H_S) + \eta^2\mathbb{E}_{\mathbf{x} \sim S}H_{\mathbf{x}}\rho H_{\mathbf{x}} - (\eta^2 R + (I - \eta H_S))\rho(\eta^2 R + (I - \eta H_S))\|_1 \\ &= \eta^2\|\mathbb{E}_{\mathbf{x} \sim S}H_{\mathbf{x}}\rho H_{\mathbf{x}} - H_S\rho H_S - \eta^2 R\rho R - (R\rho(I - \eta H_S) + (I - \eta H_S)\rho R)\|_1 \\ &\leq \eta^2(1 + 1 + \eta^2/4 + 1) < 4\eta^2. \end{aligned} \quad \square$$

242 Up to the second order term, (S10) is the imaginary time evolution of ρ under H_S . Suppose the initial state is σ , then the
 243 averaged state after T epochs is $\rho = e^{-\eta T H_S}\sigma e^{-\eta T H_S}$. Let $\beta = \eta T$ be the summation of learning rates. We can approximate
 244 the success probability of possibility by $\text{tr}(e^{-\beta H_S}\sigma e^{-\beta H_S})$ and the loss by $\text{tr}\left(H_S \frac{e^{-\beta H_S}\sigma e^{-\beta H_S}}{\text{tr}(e^{-\beta H_S}\sigma e^{-\beta H_S})}\right)$. We summarize and
 245 prove the results in the following theorem.

246 **Theorem S1.** Suppose we train the QAL model with initial state σ for T steps, with learning rate η_t at step t . Define

$$\beta = \sum_{t=1}^T \eta_t, \quad \gamma = \sum_{t=1}^T \eta_t^2, \quad \sigma(\beta) = e^{-\beta H_S}\sigma e^{-\beta H_S}. \quad (\text{S12})$$

247 Averaging over choice of training samples, the success probability of post-selection is

$$\text{tr}(\sigma(\beta)) + c_1\gamma, \quad (\text{S13})$$

248 and the average loss conditioned on the success of post-selection is

$$\frac{\text{tr}(H_S\sigma(\beta)) + c_2\gamma}{\text{tr}(\sigma(\beta)) + c_1\gamma}. \quad (\text{S14})$$

249 Here c_1, c_2 are two real numbers such that $|c_1|, |c_2| \leq 4$.

250 *Proof.* Let $\mathbf{x}_1, \dots, \mathbf{x}_T \sim S$ be the training samples in the T steps. We abbreviate $\mathbf{x}_1, \dots, \mathbf{x}_t$ as $\mathbf{x}_{1:t}$. By (S9), the unnormalized
 251 state after step t is

$$\rho_t^{\mathbf{x}_{1:t}} = (I - \eta_t H_{\mathbf{x}_t}) \cdots (I - \eta_1 H_{\mathbf{x}_1}) \sigma (I - \eta_1 H_{\mathbf{x}_1}) \cdots (I - \eta_t H_{\mathbf{x}_t}). \quad (\text{S15})$$

252 Given samples $\mathbf{x}_{1:T}$, $\text{tr}(\rho_T^{\mathbf{x}_{1:T}})$ is the success probability of post-selection and $\text{tr}\left(H_S \frac{\rho_T^{\mathbf{x}_{1:T}}}{\text{tr}(\rho_T^{\mathbf{x}_{1:T}})}\right)$ is the loss conditioned on the
 253 success of post-selection. We now average over the choice of samples. Recursively apply Lemma S2 and Lemma S1, we have

$$\mathbb{E}_{\mathbf{x}_{1:T} \sim S^T} \rho_T^{\mathbf{x}_{1:T}} = \sigma(\beta) + \gamma O, \quad (\text{S16})$$

254 for some Hermitian O with trace norm at most 4. The average success probability of post-selection is

$$\mathbb{E}_{\mathbf{x}_{1:T} \sim S^T} \text{tr}(\rho_T^{\mathbf{x}_{1:T}}) = \text{tr}(\sigma(\beta)) + c_1\gamma, \quad (\text{S17})$$

255 where $c_1 = \text{tr}(O)$ satisfies $|c_1| \leq 4$. Now we calculate the averaged loss conditioned on the success of post-selection. For
 256 clarity, denote $q = \text{tr}(\sigma(\beta)) + c_1\gamma$, $p = 1/|S|^T$ be the probability of sampling $\mathbf{x}_1, \dots, \mathbf{x}_T$. Then conditioned on the success
 257 of post-selection, the conditional probability of sampling $\mathbf{x}_{1:T}$ is $p \text{tr}(\rho_T^{\mathbf{x}_{1:T}})/q$. Therefore, the average loss conditioned on the
 258 success of post-selection is

$$\sum_{\mathbf{x}_{1:T} \sim S^T} \frac{p \text{tr}(\rho_T^{\mathbf{x}_{1:T}})}{q} \text{tr}\left(H_S \frac{\rho_T^{\mathbf{x}_{1:T}}}{\text{tr}(\rho_T^{\mathbf{x}_{1:T}})}\right) = \frac{1}{q} \mathbb{E}_{\mathbf{x}_{1:T} \sim S^T} \text{tr}(H_S \rho_T^{\mathbf{x}_{1:T}}) = \frac{\text{tr}(H_S \sigma(\beta)) + c_2\gamma}{\text{tr}(\sigma(\beta)) + c_1\gamma}, \quad (\text{S18})$$

259 where $c_2 = \text{tr}(H_S O)$ satisfies $|c_2| \leq 4$. □

260 According to the theorem, up to the second order term $c_1\gamma, c_2\gamma$, the training process behaves the same as the imaginary
 261 time evolution of σ under H_S . The effect of imaginary time evolution is clearer in the eigenbasis of H_S . Write the spectrum
 262 decomposition of H_S as $H_S = \sum_i E_i |E_i\rangle\langle E_i|$ and define $\sigma_i = \langle E_i | \sigma | E_i \rangle$ as the overlap of σ with the i -th eigenstate. Then

$$\sigma(\beta) = \sum_i \sigma_i e^{-2\beta E_i} |E_i\rangle\langle E_i|, \quad \frac{\text{tr}(H_S \sigma(\beta))}{\text{tr}(\sigma(\beta))} = \frac{\sum_i E_i \sigma_i e^{-2\beta E_i}}{\sum_i \sigma_i e^{-2\beta E_i}}. \quad (\text{S19})$$

The weight of $|E_i\rangle\langle E_i|, \sigma_i e^{-2\beta E_i}$, decays exponentially with β . The decay is slower for lower energy eigenstates. Assume σ has a non-zero overlap with the ground space. As β goes up, eventually the weight of the ground space dominates, so $\rho(\beta)/\text{Tr}(\rho(\beta))$ converges to a ground state of H_S and the empirical risk converges to the global minimum. In the following, we will make this intuition rigorous in the presence of $c_1\gamma, c_2\gamma$.

B. Convergence to global minimum

Denote the ground energy of H_S (i.e., the global minimum of the loss) as g , the projector to the ground space as Π_g , and the gap between the ground energy and the first excited state as $\delta > 0$.

Theorem S2. Suppose σ has a nonzero overlap with the ground space of H_S (that is, $\sigma_g = \text{tr}(\Pi_g \sigma) > 0$). For any constant $c \in (0, 1)$, we can choose an appropriate η and T such that if we train the QAL model for T steps with learning rate η in each step, the averaged loss conditioned on the success of post-selection is at most $g + c$.

Proof. According to Theorem S1, we only need to upper bound (S14) for $\beta = \eta T$ and $\gamma = \eta^2 T = \beta \eta$. Since

$$\begin{aligned} \frac{\text{tr}(H_S \sigma(\beta)) + c_2 \beta \eta}{\text{tr}(\sigma(\beta)) + c_1 \beta \eta} &\leq \frac{\sigma_g e^{-2\beta g} g + (1 - \sigma_g) e^{-2\beta(g+\delta)} + 4\beta \eta}{\sigma_g e^{-2\beta g} - 4\beta \eta} \\ &= g + \frac{(1 - \sigma_g) e^{-2\beta(g+\delta)} + (4 + 4g)\beta \eta}{\sigma_g e^{-2\beta g} - 4\beta \eta} \\ &\leq g + \frac{e^{-2\beta \delta} + 8\beta \eta e^{2\beta g}}{\sigma_g - 4\beta \eta e^{2\beta g}}. \end{aligned} \quad (\text{S20})$$

Choose β such that $e^{-2\beta \delta} < \sigma_g c/4$, and then choose η such that $\beta \eta e^{2\beta g} < \sigma_g c/16 < \sigma_g/16$. Then the right hand side of (S20) is at most

$$g + \frac{\sigma_g c/4 + \sigma_g c/2}{\sigma_g - \sigma_g/4} = g + c. \quad (\text{S21})$$

□

A randomly initialized state σ has a nonzero overlap with the ground space with probability 1. According to the theorem, the QAL model will converge to the global minimum of the loss function. However, this convergence is built on the success of post-selection, whose probability exponentially decays with the number of steps. Therefore, a more realistic question is whether we can build a reasonable trade-off between the success probability and the performance of the final model.

C. Convergence with constant probability

In this subsection, we will establish a practical trade-off between the accuracy of the final model and the success probability of post-selection when the initial state has a large overlap with the low-energy eigenspace of H_S .

Definition S2. Let H be a Hamiltonian. The E low energy subspace of H is the subspace spanned by the eigenstates of H with energy at most E . Denote the projector to the E low energy subspace as Π_E^H . The overlap of a state σ with the E low energy subspace is defined as $\text{tr}(\Pi_E^H \sigma)$.

Throughout this section, we focus on the Hamiltonian H_S and omit the superscript H .

Theorem S3. Let $c_1, c_3 \in (0, 1), c_2 \in (0, 1/10)$ be three constants, g be the ground energy of H_S , and $\epsilon > 0$ such that $g/\epsilon \leq c_1$. Assume the overlap between the initial state σ and the $(g + \epsilon)$ low energy eigenspace of H_S , namely $\text{tr}(\sigma \Pi_{g+\epsilon})$, is at least c_2 . Then we can choose an appropriate η and T such that if we train the QAL model with the initial state σ for T steps with learning rate η in each step, the success probability of post-selection is at least c_4 and the averaged loss conditioned on the success of post-selection is at most $g + \epsilon + c_3$. Here c_4 is a constant that only depends on c_1, c_2, c_3 .

Proof. By Theorem S1, we only need to lower bound the success probability in (S13) and the conditional loss in (S14). We will follow the notation in Theorem S1, so that $\beta = \eta T$, $\gamma = \eta^2 T$, and $\sigma(\beta) = e^{-\beta H} \sigma e^{-\beta H}$. Here we write $H = H_S$ for simplicity. We will prove the theorem for $\beta = 3 \ln(1/c_2)/(c_3 \epsilon)$, $c_4 = e^{-6(1+c_1) \ln(1/c_2)/c_3} c_2/2$ and $\gamma = c_3 c_4/40$. Accordingly, $\eta = \gamma/\beta$

295 is of order ϵ and $T = \beta^2/\gamma$ is of order $1/\epsilon^2$. By (S13), the success probability is at least

$$\begin{aligned}
 \text{tr}(\sigma(\beta)) - 4\gamma &\geq \text{tr}(\Pi_{g+\epsilon}\sigma(\beta)) - 4\gamma \\
 &= \text{tr}(e^{-\beta H}\Pi_{g+\epsilon}e^{-\beta H}\sigma) - 4\gamma \\
 &\geq e^{-2\beta(g+\epsilon)} \text{tr}(\Pi_{g+\epsilon}\sigma) - 4\gamma \\
 &\geq e^{-2\beta\epsilon(1+c_1)}c_2 - 4\gamma \\
 &= 2c_4 - 4\gamma > c_4.
 \end{aligned} \tag{S22}$$

296 By (S14), the averaged loss conditioned on the success of post-selection is at most

$$\begin{aligned}
 \frac{\text{tr}(H\sigma(\beta)) + 4\gamma}{\text{tr}(\sigma(\beta)) - 4\gamma} &= g + \frac{\text{tr}((H - gI)\sigma(\beta))}{\text{tr}(\sigma(\beta)) - 4\gamma} + \frac{(4 + 4g)\gamma}{\text{tr}(\sigma(\beta)) - 4\gamma} \\
 &\leq g + \frac{\text{tr}((H - gI)\sigma(\beta))}{\text{tr}(\sigma(\beta))(1 - c_3/20)} + \frac{8\gamma}{c_4} \\
 &\leq g + \frac{c_3}{5} + (1 + \frac{c_3}{5}) \frac{\text{tr}((H - gI)\sigma(\beta))}{\text{tr}(\sigma(\beta))},
 \end{aligned} \tag{S23}$$

297 where we use $4\gamma = c_3c_4/10 \leq c_3 \text{tr}(\sigma(\beta))/20$ in the second line and $(1 + c_3/5)(1 - c_3/20) \geq 1$ in the third line. So it
 298 suffices to upper bound $\text{tr}((H - gI)\sigma(\beta))/\text{tr}(\sigma(\beta))$. Write the spectrum decomposition of H as $H = \sum_i E_i |E_i\rangle\langle E_i|$ and let
 299 $x_i = 2\beta(E_i - g)$, $\sigma_i = \langle E_i|\sigma|E_i\rangle$. We simplify the last term of (S23) to

$$\begin{aligned}
 \frac{\text{tr}((H - gI)\sigma(\beta))}{\text{tr}(\sigma(\beta))} &= \frac{\sum_i (E_i - g)e^{-2\beta E_i} \sigma_i}{\sum_i e^{-2\beta E_i} \sigma_i} \\
 &= \frac{\sum_i \sigma_i e^{-x_i} x_i}{\sum_i \sigma_i e^{-x_i}} \cdot \frac{1}{2\beta}.
 \end{aligned} \tag{S24}$$

300 Split the Hilbert space into low energy and high energy eigenspaces, $I_L = \{i : E_i \leq g + \epsilon\}$ and $I_H = \{i : E_i > g + \epsilon\}$. Let
 301 $p_L = \text{tr}(\sigma\Pi_{g+\epsilon}) = \sum_{i \in I_L} \sigma_i \geq c_2$ and $p_H = 1 - p_L$ be the overlaps of σ with the two subspaces. Since $f(y) = -y \ln(y)$ is
 302 concave, by Jensen's inequality,

$$\sum_{i \in L} \frac{\sigma_i}{p_L} f(e^{-x_i}) \leq f\left(\sum_{i \in L} \frac{\sigma_i}{p_L} e^{-x_i}\right). \tag{S25}$$

303 Let l be the number such that $e^{-l} = \sum_{i \in L} \sigma_i e^{-x_i}/p_L$. The inequality becomes $\sum_{i \in L} \sigma_i e^{-x_i} x_i \leq p_L e^{-l} l$. Similarly, let
 304 $h = -\ln(\sum_{i \in H} \sigma_i e^{-x_i}/\sum_{i \in H} \sigma_i)$, then $\sum_{i \in H} \sigma_i e^{-x_i} x_i \leq p_H e^{-h} h$. Therefore,

$$\frac{\sum_i \sigma_i e^{-x_i} x_i}{\sum_i \sigma_i e^{-x_i}} \leq \frac{p_L e^{-l} l + p_H e^{-h} h}{p_L e^{-l} + p_H e^{-h}}. \tag{S26}$$

305 By definition, $x_i \in [0, 2\beta\epsilon]$ for $i \in L$ and $x_i > 2\beta\epsilon$ for $i \in H$. Since e^{-l} is a mixed of e^{-x_i} ($i \in L$), we have $0 \leq l \leq 2\beta\epsilon$ and
 306 similarly $h \geq 2\beta\epsilon$. Denote $y = h - l$. Insert (S24) and (S26) to (S23).

$$\begin{aligned}
 \frac{\text{tr}(H\sigma(\beta)) + 4\gamma}{\text{tr}(\sigma(\beta)) - 4\gamma} &\leq g + \frac{c_3}{5} + (1 + \frac{c_3}{5}) \frac{p_L e^{-l} l + p_H e^{-h} h}{p_L e^{-l} + p_H e^{-h}} \cdot \frac{1}{2\beta} \\
 &= g + \frac{c_3}{5} + (1 + \frac{c_3}{5}) (l + \frac{p_H e^{-y} y}{p_L + p_H e^{-y}}) \cdot \frac{1}{2\beta} \\
 &\leq g + \frac{c_3}{5} + (1 + \frac{c_3}{5}) (l + \frac{e^{-y} y}{c_2 + e^{-y}}) \cdot \frac{1}{2\beta}.
 \end{aligned} \tag{S27}$$

307 By differentiating $g(y) = e^{-y}y/(c_2 + e^{-y}) = y/(c_2 e^y + 1)$, we find that $g(y) \leq g(y^*)$ for the $y^* > 0$ such that $c_2 e^{y^*} (y^* - 1) = 1$.
 308 For this y^* we have $g(y^*) = y^* - 1$. Assume $y^* > \ln(1/c_2) > 2$, then $1 = c_2 e^{y^*} (y^* - 1) > c_2 (1/c_2) (2 - 1) = 1$, a contradiction.
 309 So $g(y) \leq g(y^*) = y^* - 1 \leq \ln(1/c_2)$. By (S27), the averaged loss conditioned on the success of post-selection is at most

$$g + \frac{c_3}{5} + (1 + \frac{c_3}{5}) (2\beta\epsilon + \ln(1/c_2)) \cdot \frac{1}{2\beta} = g + \frac{c_3}{5} + (1 + \frac{c_3}{5}) (\epsilon + \frac{c_3\epsilon}{6}) < g + \epsilon + c_3. \tag{S28}$$

□

310 The theorem ensures the convergence of the QAL training process with a constant success probability, assuming a good initial
 311 state. Concretely, as long as the initial state has a large (at least c_2) overlap with the low energy eigenspace (with energy at most
 312 $g + \epsilon$), the QAL training process will converge to the low energy eigenspace (up to an arbitrarily small residue error c_3) with a
 313 constant success probability (c_4 that only depends on c_1, c_2, c_3).

D. Heavy-tailed Hamiltonian

Theorem S3 highlights the importance of the initial state σ . However, without prior knowledge of H_S , we cannot do better than a random guess, or equivalently, starting from the maximally-mixed state $\sigma = I/2^n$. Therefore, we actually hope that H_S has a constant proportion of low energy eigenstates that do not scale up with n , the dimension of \mathbf{x} , and the size m of the training dataset. We formalize this intuition in the following definition.

Definition S3 (Heavy-tailed Hamiltonian). *We say a Hamiltonian H is (E, c) -heavy-tailed if the proportion of eigenstates with energy at most E is at least c .*

Theorem S4. *Let $c_1, c_3 \in (0, 1)$, $c_2 \in (0, 1/10)$ be three constants. Suppose H_S is $(g + \epsilon, c_2)$ -heavy-tailed, where g is the ground energy of H_S and $\epsilon > 0$ such that $g/\epsilon \leq c_1$. Then we can choose an appropriate η and T such that if we train the QAL model with a maximally-mixed initial state in computational basis for T steps with learning rate η in each step, the success probability of post-selection is at least c_4 and the averaged loss conditioned on the success of post-selection is at most $g + \epsilon + c_3$. Here c_4 is a constant that only depends on c_1, c_2, c_3 .*

Proof. By definition of heavy-tailed Hamiltonian, the overlap between the initial state $\sigma = I/2^n$ and the $(g + \epsilon)$ low eigenspace of H_S is at least c_2 . The theorem follows directly from Theorem S3. \square

Therefore, the QAL training process is guaranteed to converge to the low energy eigenspace of a heavy-tailed H_S . We now argue that when H_S comes from a reasonable dataset, it is likely to be heavy-tailed due to the similarity of data. Consider the extremely simple example of classifying dogs and cats, where all dogs look similar and all cats look similar. The Hamiltonian H_S is approximately a mixture of two projectors of dimensions 2^{n-1} , H_{dogs} and H_{cats} . Regard H_{dogs} and H_{cats} as random projectors, then H_S has a constant proportion of near-zero eigenvalues. This assumption is supported by the numerical simulation.

Remark that while Theorem S4 applies to the maximally-mixed initial state, in reality we will use a random initial state in the computational basis. Once we sample an initial state better than the maximally-mixed state, we can stick to it and apply Theorem S3. Furthermore, the QAL framework can be combined with the flipped model defined in Definition S1 to achieve a favorable tradeoff between the number of circuit runs and the post-selection probability. Specifically, we first select a shallow ansatz $V(\theta)$ and train the flipped model as defined in Definition S1. The trained state $V(\theta)|0\rangle$ is then used as the initial state of the QAL model, which is expected to exhibit a much larger overlap with the low-energy subspace than a random initialization.

A possible way to relax the heavy-tail assumption is to adopt alternative ground-state preparation methods beyond imaginary-time evolution. For instance, Ref. [27] introduces a ground-state preparation algorithm based on real-time quantum dynamics simulation. Unlike projection-based methods, their approach does not rely on an initial state with substantial overlap with the low-energy subspace. Investigating whether such a technique can enhance the efficiency of the QAL scheme constitutes an interesting future direction. Nevertheless, we emphasize that any feasible ground-state preparation method must inherently rely on certain structural assumptions about the data Hamiltonian, as the general ground-state preparation problem is known to be computationally hard.

IV. OTHER PROPERTIES OF QUANTUM AUTOMATED LEARNING

A. Generalization

The previous results establish the explainable trainability of QAL. While in training classical neural networks, each epoch is a full pass of the training dataset, in QAL, each step only involves a single datum. This indicates that the QAL model could be optimized using a few data points. In this subsection we rigorously demonstrate the generalization ability of QAL, showing that once the model achieves a good performance on a small dataset, the good performance will generalize to unseen data.

Recall that the training loss and the true loss of $|\psi\rangle$ are $\hat{R}_S(\psi) = \mathbb{E}_{\mathbf{x} \sim S} \langle \psi | H_{\mathbf{x}} | \psi \rangle$ and $R(\psi) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \langle \psi | H_{\mathbf{x}} | \psi \rangle$, respectively.

Theorem S5. *With probability at least $1 - \delta$ over the choice of S , the generalization gap is upper bounded by*

$$\max_{|\psi\rangle} (R(\psi) - \hat{R}_S(\psi)) \leq \sqrt{\frac{4 \ln(2^{n+1}/\delta)}{m}}. \quad (\text{S29})$$

Proof. By definition, the left-hand side is upper bounded by the spectral norm of $\mathbb{E}_{\mathbf{x} \sim S} H_{\mathbf{x}} - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} H_{\mathbf{x}}$, which can be bounded by matrix Bernstein inequality (see, e.g., [28, Theorem 6.1.1]):

$$\Pr_S[\|\mathbb{E}_{\mathbf{x} \sim S} H_{\mathbf{x}} - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} H_{\mathbf{x}}\|_{\infty} \geq t] \leq 2^{n+1} \exp(-mt^2/4).$$

The theorem follows by setting $t = \sqrt{4 \ln(2^{n+1}/\delta)/m}$. \square

According to the theorem, as long as the size m of the training dataset is larger than $\Omega(n)$ (i.e., a logarithm of the degree of freedom), a model state $|\psi\rangle$ with low training loss has a low true loss with high probability. This is better than quantum neural networks where the training dataset size has to be larger than the degree of freedom [6, 7]. From the proof of the theorem, it is clear that the good generalization stems from the simple quadratic form of the loss function.

B. Universal representation power

We have shown that the simple quadratic form of the loss function endows QAL with good generalization ability. One may naturally ask whether this simplicity compromises the expressive power of the model. We show that this is not the case by proving that the QAL protocol possesses universal representation power. Here, the universal representation property refers to the ability to approximate any continuous function [29, 30], which has been established for quantum machine learning models in Ref. [31].

Lemma S3 (Result 1 of Ref. [31]). *Let \mathcal{X} be a compact subset of \mathbb{R}^l representing the feasible data domain. For any continuous function $g : \mathcal{X} \rightarrow \mathbb{R}$ and any $\epsilon > 0$, there exist a number of qubits n_0 , an n_0 -qubit single-layer parametrized quantum circuit $V_{n_0}(\mathbf{x})$, and a Hermitian observable O , such that*

$$|\langle 0|^{\otimes n_0} V_{n_0}^\dagger(\mathbf{x}) O V_{n_0}(\mathbf{x}) |0\rangle^{\otimes n_0} - g(\mathbf{x})| < \epsilon, \quad \forall \mathbf{x} \in \mathcal{X}. \quad (\text{S30})$$

Theorem S6 (Universal representation power of QAL). *Let $\mathcal{K}_0, \dots, \mathcal{K}_{k-1} \subset \mathbb{R}^l$ be disjoint finite closed sets. For any $\eta \in (0, 1)$, there exist a number of qubits n , a model state $|\psi\rangle$, and an encoding unitary $U(\mathbf{x})$ such that, for every $\mathbf{x} \in \mathcal{K}_y$, measuring the middle $\kappa = \lceil \log k \rceil$ qubits of $U(\mathbf{x}) |\psi\rangle$ yields the correct label y with probability at least $1 - \eta$.*

Proof. For simplicity we assume k is a power of two, i.e., $k = 2^\kappa$. Otherwise, we embed the labels into $\{0, \dots, 2^\kappa - 1\}$ and ignore unused outcomes. By definition, $\mathcal{K} = \bigcup_{i=0}^{k-1} \mathcal{K}_i$ is a compact subset of \mathbb{R}^l . It is easy to see that there exists a continuous function $g : \mathcal{K} \rightarrow [0, k]$ such that $g(\mathbf{x}) \in [y - \frac{1}{8}, y + \frac{1}{8}]$, $\forall \mathbf{x} \in \mathcal{K}_y$.

By Lemma S3 with accuracy $\epsilon_0 = \frac{1}{8}$, there exist n_0 , a variational circuit $U_0(\mathbf{x})$, and a Hermitian operator O such that

$$\langle 0|^{\otimes n_0} U_0^\dagger(\mathbf{x}) O U_0(\mathbf{x}) |0\rangle^{\otimes n_0} \in [y - \frac{1}{4}, y + \frac{1}{4}], \quad \forall \mathbf{x} \in \mathcal{K}_y. \quad (\text{S31})$$

Without loss of generality, O can be chosen as a diagonal observable $\sum_i \lambda_i |i\rangle\langle i|$ in the computational basis, by absorbing any additional unitary transformation into $U_0(\mathbf{x})$.

Fix an integer $b \geq 1$ (to be chosen later). Consider a larger system with $n = n_0 b + \kappa$ qubits. We partition the Hilbert space into orthogonal subspaces

$$A_y = \text{span} \left(\{ |i_1 i_2 \dots i_b 0^\kappa\rangle \mid y - \frac{1}{2} \leq \frac{\lambda_{i_1} + \dots + \lambda_{i_b}}{b} < y + \frac{1}{2} \} \cup B_y \right), \quad (\text{S32})$$

where B_y contains some other computational basis states to ensure that the dimension of each A_y is $2^n/2^\kappa$. Let Π_y be the projector onto subspace A_y . Then $\{\Pi_y\}_y$ forms a PVM on n qubits.

Let $U_1(\mathbf{x}) = U_0(\mathbf{x})^{\otimes b} \otimes \mathbf{I}^{\otimes \kappa}$ and set the model state $|\psi\rangle = |0\rangle^{\otimes n}$. We now prove that, if we measure $U_1(\mathbf{x}) |\psi\rangle$ for $\mathbf{x} \in \mathcal{K}_y$ using PVM $\{\Pi_{y'}\}_{y'}$, the outcome is y with high probability. Indeed, the probability of outcome y is

$$\begin{aligned} & \langle \psi | U_1^\dagger(\mathbf{x}) \Pi_y U_1(\mathbf{x}) | \psi \rangle \\ & \geq \sum_{i_1, \dots, i_b} \langle \psi | U_0^\dagger(\mathbf{x}) |i_1 i_2 \dots i_b 0^\kappa\rangle \langle i_1 i_2 \dots i_b 0^\kappa | U_0(\mathbf{x}) | \psi \rangle \cdot \mathbf{1}[y - \frac{1}{2} \leq \frac{\lambda_{i_1} + \dots + \lambda_{i_b}}{b} < y + \frac{1}{2}] \\ & = \sum_{i_1, \dots, i_b} \mathbf{1}[y - \frac{1}{2} \leq \frac{\lambda_{i_1} + \dots + \lambda_{i_b}}{b} < y + \frac{1}{2}] \cdot \prod_{j=1}^b \left| \langle 0^{n_0} | U_0^\dagger(\mathbf{x}) | i_j \rangle \right|^2. \end{aligned} \quad (\text{S33})$$

Here $\mathbf{1}[E]$ is the indicator function that takes value 1 if the event E is true and 0 otherwise. The last line can be interpreted as the probability of event $y - \frac{1}{2} \leq \frac{\lambda_{i_1} + \dots + \lambda_{i_b}}{b} < y + \frac{1}{2}$, where i_1, \dots, i_b are independently sampled from the distribution $|\langle 0^{n_0} | U_0^\dagger(\mathbf{x}) | i \rangle|^2$. By (S31), $\sum_i \lambda_i |\langle 0^{n_0} | U_0^\dagger(\mathbf{x}) | i \rangle|^2 \in [y - \frac{1}{4}, y + \frac{1}{4}]$, which means the expectation of each λ_i (denoted by $\bar{\lambda}$) is in $[y - \frac{1}{4}, y + \frac{1}{4}]$. By Hoeffding's inequality of concentration,

$$\Pr \left[\left| \frac{\lambda_{i_1} + \dots + \lambda_{i_b}}{b} - \bar{\lambda} \right| < \frac{1}{4} \right] \geq 1 - 2 \exp \left(-\frac{b}{8R^2} \right), \quad (\text{S34})$$

where $R = \max_i \{\lambda_i\} - \min_i \{\lambda_i\}$. Setting $b = 8R^2 \ln(2/\eta)$, we obtain that $\langle \psi | U_1^\dagger(\mathbf{x}) \Pi_y U_1(\mathbf{x}) | \psi \rangle \geq 1 - \eta$.

We have proved that, if we measure $U_1(\mathbf{x}) | \psi \rangle$ for $\mathbf{x} \in \mathcal{K}_y$ using PVM $\{\Pi_{y'}\}_{y'}$, the outcome is y with probability at least $1 - \eta$. Since the dimension of each $\Pi_{y'}$ is $2^n/2^\kappa$, there exists a unitary U_2 such that the PVM $\{U_2^\dagger \Pi_{y'} U_2\}_{y'}$ is the computational basis measurement on the middle κ qubits. The theorem follows by setting $U(\mathbf{x}) = U_2 U_1(\mathbf{x})$. \square

C. State reusability

Suppose we have a well-trained model state $|\psi\rangle$. To predict the label of a new data sample \mathbf{x} , we apply $U(\mathbf{x})$ to $|\psi\rangle$ and measure the middle qubits. The measurement outcome yields the correct label $y(\mathbf{x})$ with high probability. However, since quantum measurements are destructive, one might worry that the model state is destroyed after a single prediction, requiring retraining for each new data point. Here we argue that this problem is not as severe as it may appear.

In Ref. [32], Aaronson introduced the task of *shadow tomography*: given M observables $0 \leq O_1, \dots, O_M \leq \mathbf{I}$ and copies of an unknown quantum state ρ , the goal is to estimate all expectation values $\text{Tr}(\rho O_1), \dots, \text{Tr}(\rho O_M)$. Remarkably, he showed that only $\text{polylog}(M)$ copies of ρ suffice to estimate all these quantities within small additive error. We cite below an improved version of this result in the online setting from Ref. [33].

Lemma S4 (Theorem 1.4 of [33]). *There exists a quantum algorithm that, given $M \in \mathbb{N}$, $0 < \epsilon < \frac{1}{2}$, and $\tilde{O}(n \log^2 M / \epsilon^4)$ copies of ρ , behaves as follows: when presented sequentially with any series of observables O_1, \dots, O_M satisfying $0 \leq O_i \leq \mathbf{I}$, the algorithm outputs an estimate of $\text{Tr}(\rho O_t)$ upon receiving each O_t . Except with probability at most δ , all M estimates have errors at most ϵ .*

Therefore, for M unseen data samples $\mathbf{x}_1, \dots, \mathbf{x}_M$, we can invoke shadow tomography to estimate all probabilities $\text{Tr}(|\psi\rangle\langle\psi| U(\mathbf{x}_i)^\dagger \Pi_y U(\mathbf{x}_i))$ for $i \leq M$ and $y \leq k$, within small constant error, using only $\tilde{O}(n \log^2 M)$ copies of $|\psi\rangle\langle\psi|$. The resulting estimates suffice to predict the labels of all M data points with high accuracy. Moreover, the online feature of Lemma S4 allows the samples $\mathbf{x}_1, \dots, \mathbf{x}_M$ to arrive sequentially, rather than being known in advance, which is more consistent with typical machine learning settings. However, although shadow tomography is extremely efficient in sample complexity, the known algorithms require exponential computation time. Hence, it serves mainly as an information-theoretic argument rather than a practical approach to reusability.

We next present a more practical perspective on state reusability based on the notion of *gentle measurements*. The intuition is that if the measurement outcome is almost deterministic, the measurement only weakly disturbs the state. For instance, measuring the state $\epsilon|0\rangle + \sqrt{1-\epsilon^2}|1\rangle$ in the computational basis barely changes the state when $\epsilon \ll 1$. This idea is captured formally by the following result from Gao [34].

Lemma S5 (Gentle measurement lemma). *Let ρ be a quantum state and P_1, \dots, P_M a sequence of projectors such that $\text{Tr}(\rho P_i) = 1 - \epsilon_i$ for $i = 1, \dots, M$ (we call $1 - \epsilon_i$ the gentleness of P_i). Suppose we measure ρ sequentially with $\{P_1, \mathbf{I} - P_1\}, \{P_2, \mathbf{I} - P_2\}, \dots, \{P_M, \mathbf{I} - P_M\}$. Then:*

1. *The probability of obtaining the sequence of outcomes P_1, \dots, P_M is at least $1 - 4 \sum_{i=1}^M \epsilon_i$.*
2. *Conditioned on obtaining these outcomes, the trace distance between the post-measurement state and the original state ρ is at most $2\sqrt{\sum_{i=1}^M \epsilon_i}$.*

In our case, assume that $|\psi\rangle$ is well-trained so that for each of M unseen data samples $\mathbf{x}_1, \dots, \mathbf{x}_M$, we have

$$\langle \psi | U(\mathbf{x}_i)^\dagger \Pi_{y(\mathbf{x}_i)} U(\mathbf{x}_i) | \psi \rangle \geq 1 - \epsilon.$$

Let $\kappa = \lceil \log k \rceil$. The computational-basis measurement on the middle κ qubits can be decomposed into κ single-qubit two-outcome measurements, each with gentleness at least $1 - \epsilon$. To predict the labels of $\mathbf{x}_1, \dots, \mathbf{x}_M$, we thus perform $M\kappa$ such measurements. By Lemma S5, the probability that all these measurements yield the desired outcomes (i.e., all predictions are correct) is at least $1 - 4M\kappa\epsilon$. Furthermore, conditioned on all predictions being correct, the post-measurement state deviates from the original $|\psi\rangle$ by at most $2\sqrt{M\kappa\epsilon}$ in trace distance. Hence, the model's performance can be recovered by a few additional training steps, provided that $M\kappa\epsilon \ll 1$.

The argument above applies when $M \ll 1/(\kappa\epsilon)$. This limit can be exponentially improved via a *majority-vote amplification* technique if we are allowed to perform joint measurements on multiple copies of $|\psi\rangle$. For an odd integer $r \in \mathbb{N}$, let C_r^{maj} denote a reversible classical circuit that maps a bit string $z \in \{0, 1\}^r$ to its majority value (output on the middle wire). By quantizing each gate in C_r^{maj} , we obtain a unitary U_r^{maj} of dimension $2^r \times 2^r$ whose middle qubit encodes the majority value of the input computational basis string.

Lemma S6 (Majority-vote amplification). Let $r \in \mathbb{N}$ and $\epsilon \in (0, \frac{1}{4})$. Let A be the set of measured qubits in the QAL model and \bar{A} its complement, so that $|A| = \kappa$. Suppose that when measuring the qubits in A of $|\psi(\mathbf{x})\rangle = U(\mathbf{x})|\psi\rangle$, the outcome yields $y(\mathbf{x})$ with probability at least $1 - \epsilon$. Consider the state $(U_r^{\text{maj}})^{\otimes A} |\psi(\mathbf{x})\rangle^{\otimes r}$, i.e., r registers each in state $|\psi(\mathbf{x})\rangle$, followed by κ unitaries U_r^{maj} acting respectively on the a -th qubits across all registers for $a \in A$. Then, measuring the qubits in A of the middle register of $(U_r^{\text{maj}})^{\otimes A} |\psi(\mathbf{x})\rangle^{\otimes r}$ yields $y(\mathbf{x})$ with probability at least $1 - \kappa(4\epsilon(1 - \epsilon))^{r/2}$.

Proof. Fix any $i \leq \kappa$. The probability that the i -th output bit coincides with the i -th bit of $y(\mathbf{x})$ equals the probability that the majority of r independent Bernoulli trials with success probability $1 - \epsilon$ yield a success. By the Chernoff bound, this probability is at least $1 - (4\epsilon(1 - \epsilon))^{r/2}$. The lemma follows by applying a union bound over all κ bits. \square

Therefore, by performing majority voting across r copies of $|\psi\rangle$, we can boost the prediction accuracy from $1 - \epsilon$ to $1 - \kappa(4\epsilon(1 - \epsilon))^{r/2}$. Combining this result with Lemma S5, we arrive at the following reusability result.

Theorem S7 (Reusability of QAL model states). Let $\mathbf{x}_1, \dots, \mathbf{x}_M$ be M unseen data, $\epsilon > 0$. Suppose the QAL training process yields a good model state $|\psi\rangle$ such that $\langle \psi | U(\mathbf{x}_i)^\dagger \Pi_{y(\mathbf{x}_i)} U(\mathbf{x}_i) | \psi \rangle \geq 1 - \epsilon$, $\forall i = 1, \dots, M$. For an odd number $r > 0$, if we train r copies of $|\psi\rangle$ and predict $\mathbf{x}_1, \dots, \mathbf{x}_M$ sequentially using majority-vote amplification, we correctly predict all labels with probability at least $1 - 4M\kappa(4\epsilon(1 - \epsilon))^{r/2}$. Furthermore, conditioned on all correct labels, the post-measurement state only deviate from $|\psi\rangle^{\otimes r}$ by at most $2\sqrt{M\kappa(4\epsilon(1 - \epsilon))^{r/2}}$ in trace distance.

D. Mini-batch optimization

In QAL training, each optimization step involves only a single data sample, which appears incompatible with batch or mini-batch training strategies. As a result, traversing the entire dataset requires at least as many training steps as the dataset size, which may lead to prohibitively deep training circuits.

Here we describe a method that allows us to exploit information from all training data while keeping the circuit depth shallow. The idea is to employ the majority-vote mechanism discussed in Lemma S6, but now the r copies of the model state, $|\psi_1\rangle, \dots, |\psi_r\rangle$, are independently trained on different subsets of the training data, ensuring that each data sample participates in at least one of the trainings. If we view the overall circuit as a single QAL model acting on rn qubits, this ensemble-based construction effectively approximates the behavior of mini-batch optimization, allowing the model to aggregate information from multiple data points in parallel without increasing the depth.

-
- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).
 - [2] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, *Nat. Rev. Phys.* **3**, 625 (2021).
 - [3] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, *Phys. Rev. A* **98**, 032309 (2018).
 - [4] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, *Phys. Rev. A* **99**, 032331 (2019).
 - [5] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum Natural Gradient, *Quantum* **4**, 269 (2020).
 - [6] H. Cai, Q. Ye, and D.-L. Deng, Sample complexity of learning parametric quantum circuits, *Quantum Sci. Technol.* **7**, 025014 (2022).
 - [7] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Generalization in quantum machine learning from few training data, *Nat. Commun.* **13**, 4919 (2022).
 - [8] E. Recio-Armengol, S. Ahmed, and J. Bowles, Train on classical, deploy on quantum: Scaling generative quantum machine learning to a thousand qubits (2025), 2503.02934.
 - [9] A. Abbas, R. King, H.-Y. Huang, W. J. Huggins, R. Movassagh, D. Gilboa, and J. R. McClean, On quantum backpropagation, information reuse, and cheating measurement collapse, in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 (Curran Associates Inc.) pp. 44792–44819.
 - [10] X. You and X. Wu, Exponentially Many Local Minima in Quantum Neural Networks, in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021) pp. 12144–12155.
 - [11] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, *Nat. Commun.* **9**, 4812 (2018).
 - [12] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost Function Dependent Barren Plateaus in Shallow Parametrized Quantum Circuits, *Nat. Commun.* **12**, 1791 (2021).
 - [13] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement-Induced Barren Plateaus, *PRX Quantum* **2**, 040316 (2021).
 - [14] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, *PRX Quantum* **3**, 010313 (2022).
 - [15] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, *Nat. Commun.* **12**, 6961 (2021).

- [16] M. Larocca, S. Thanasilp, S. Wang, K. Sharma, J. Biamonte, P. J. Coles, L. Cincio, J. R. McClean, Z. Holmes, and M. Cerezo, A Review of Barren Plateaus in Variational Quantum Computing, [arXiv:2405.00781](#) (2024).
- [17] A. Părez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Data re-uploading for a universal quantum classifier, [Quantum](#) **4**, 226 (2020).
- [18] J. Chen, Y. Wu, Z. Yang, S. Xu, X. Ye, D. Li, K. Wang, C. Zhang, F. Jin, X. Zhu, Y. Gao, Z. Tan, Z. Cui, A. Zhang, N. Wang, Y. Zou, T. Li, F. Shen, J. Zhong, Z. Bao, Z. Zhu, Z. Song, J. Deng, H. Dong, P. Zhang, W. Zhang, H. Li, Q. Guo, Z. Wang, Y. Li, X. Wang, C. Song, and H. Wang, [Quantum ensemble learning with a programmable superconducting processor](#) (2024), 2503.11047.
- [19] S. Jerbi, C. Gyurik, S. C. Marshall, R. Molteni, and V. Dunjko, Shadows of quantum machine learning, [Nat. Commun.](#) **15**, 5676 (2024).
- [20] A. M. Childs and N. Wiebe, Hamiltonian Simulation Using Linear Combinations of Unitary Operations, [QIC](#) **12**, 901 (2012).
- [21] I. M. Georgescu, S. Ashhab, and F. Nori, Quantum simulation, [Rev. Mod. Phys.](#) **86**, 153 (2014).
- [22] G. H. Low and I. L. Chuang, Optimal hamiltonian simulation by quantum signal processing, [Phys. Rev. Lett.](#) **118**, 010501 (2017).
- [23] L. Clinton, J. Bausch, and T. Cubitt, Hamiltonian simulation algorithms for near-term quantum hardware, [Nat. Commun.](#) **12**, 4989 (2021).
- [24] S. Lloyd, M. Mohseni, and P. Rebentrost, Quantum principal component analysis, [Nat. Phys.](#) **10**, 631 (2014).
- [25] R. Vale, T. M. D. Azevedo, I. C. S. Araújo, I. F. Araujo, and A. J. da Silva, Decomposition of Multi-controlled Special Unitary Single-Qubit Gates, [arXiv:2302.06377](#) (2023).
- [26] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning* (MIT press, 2018).
- [27] D. Motlagh, M. S. Zini, J. M. Arrazola, and N. Wiebe, Ground state preparation via dynamical cooling (2024), 2404.05810.
- [28] J. A. Tropp, An Introduction to Matrix Concentration Inequalities, [arXiv:1501.01571](#) (2015).
- [29] G. Cybenko, Approximation by superpositions of a sigmoidal function, [Math. Control Signal Systems](#) **2**, 303 (1989).
- [30] K. Hornik, Approximation capabilities of multilayer feedforward networks, [Neural Networks](#) **4**, 251 (1991).
- [31] T. Goto, Q. H. Tran, and K. Nakajima, Universal Approximation Property of Quantum Machine Learning Models in Quantum-Enhanced Feature Spaces, [Phys. Rev. Lett.](#) **127**, 090506 (2021).
- [32] S. Aaronson, Shadow tomography of quantum states, in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018 (Association for Computing Machinery, New York, NY, USA, 2018) pp. 325–338.
- [33] C. Bădescu and R. O’Donnell, Improved Quantum data analysis, in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (ACM, Virtual Italy, 2021) pp. 1398–1411.
- [34] J. Gao, Quantum union bounds for sequential projective measurements, [Physical Review A](#) **92**, 052331 (2015), [arxiv:1410.5688 \[quant-ph\]](#).