# Training robust and generalizable quantum models

Julian Berberich,[1,*] Daniel Fink,[2] Daniel Pranjić,[3] Christian Tutschku,[3] and Christian Holm[2]
[1]*University of Stuttgart, Institute for Systems Theory and Automatic Control, 70569 Stuttgart, Germany*
[2]*University of Stuttgart, Institute for Computational Physics, 70569 Stuttgart, Germany*
[3]*Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering, 70569 Stuttgart, Germany*

(Received 3 May 2024; accepted 28 November 2024; published 27 December 2024)

Adversarial robustness and generalization are both crucial properties of reliable machine learning models. In this paper, we study these properties in the context of quantum machine learning based on Lipschitz bounds. We derive parameter-dependent Lipschitz bounds for quantum models with trainable encoding, showing that the norm of the data encoding has a crucial impact on the robustness against data perturbations. Further, we derive a bound on the generalization error which explicitly involves the parameters of the data encoding. Based on these theoretical results, we propose a practical strategy for training robust and generalizable quantum models by regularizing the Lipschitz bound in the cost. Moreover, we show that, for fixed and nontrainable encodings, as those frequently employed in quantum machine learning, the Lipschitz bound cannot be influenced by tuning the parameters. Thus trainable encodings are crucial for systematically adapting robustness and generalization during training. The practical implications of our theoretical findings are illustrated with numerical results.

DOI: 10.1103/PhysRevResearch.6.043326

## I. INTRODUCTION

Robustness of machine learning (ML) models is an increasingly important property, especially when operating on real-world data subject to perturbations. In practice, there are various possible sources of perturbations such as noisy data acquisition or adversarial attacks. The latter are tiny but carefully chosen manipulations of the data, and they can lead to dramatic misclassification in neural networks [1,2]. As a result, much research has been devoted to better understanding and improving adversarial robustness [3–5]. It is well-known that robustness is closely connected to generalization [1,2,6–8], i.e., the ability of a model to extrapolate beyond the training data. Intuitively, if a model is robust then small input changes only cause small output changes, thus counteracting the risk of overfitting.

A Lipschitz bound of a model $f$ is any $L > 0$ satisfying

$$\|f(x_1) - f(x_2)\| \leqslant L\|x_1 - x_2\| \tag{1}$$

for all $x_1, x_2 \in \mathcal{D} \subseteq \mathbb{R}^d$, where $d$ is the data dimension. By definition, Lipschitz bounds quantify the worst-case output change that can be caused by data perturbations and, thus, they provide a useful measure of adversarial robustness. Therefore they are a well-established tool for characterizing robustness and generalization properties of ML models [2,6,9–15]. Lipschitz bounds cannot only be used to better understand these two properties, but they also allow one to improve them by regularizing the Lipschitz bound during training [2,6,16–18].

In this paper, we study the interplay of robustness and generalization in quantum machine learning (QML). Variational quantum circuits are a well-studied class of quantum models [20–23] and they promise benefits over classical ML in various aspects including trainability, expressivity, and generalization performance [24,25]. Data reuploading circuits generalize the classical variational circuits by concatenating a data encoding and a parametrized quantum circuit not only once but repeatedly, thus iterating between data- and parameter-dependent gates [26]. This alternation provides substantial improvements on expressivity, leading to a universal quantum classifier even in the single-qubit case [26,28].

Just as in the classical case, robustness is crucial for quantum models. First, if QML is to provide benefits over classical ML, it is necessary to implement QML circuits which are robust with respect to quantum errors occurring due to imperfect hardware in the noisy intermediate-scale quantum (NISQ) era [29]. Questions of robustness of quantum models against such hardware errors have been studied, e.g., in Refs. [30,31]. Lipschitz bounds can be used to study robustness of quantum algorithms against certain types of hardware errors, e.g., coherent control errors [32].

However, robustness against hardware errors is entirely different from and independent of the robustness of a quantum model against data perturbations, which is the subject of this paper. The latter type of robustness has been studied in the context of quantum adversarial machine learning [33,34]. Not surprisingly, just like their classical counterparts, quantum models are also vulnerable to adversarial attacks, both when operating based on classical data [35,36] and quantum data [35,37–42]. To mitigate these attacks, it is desirable to design

*Contact author: julian.berberich@ist.uni-stuttgart.de

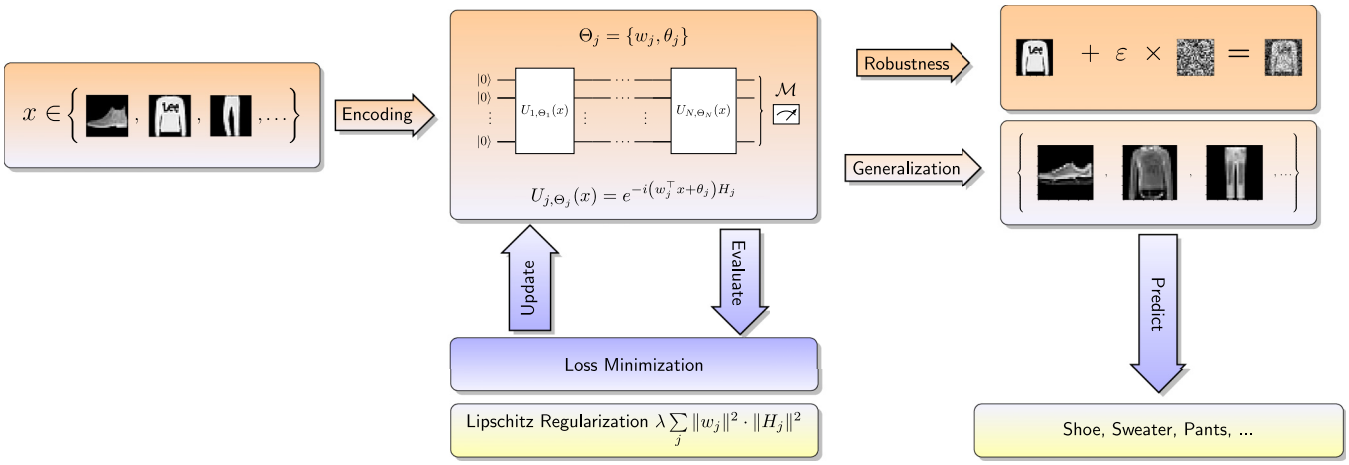


FIG. 1. Schematic illustration of the quantum model and training setup considered in this work for an exemplary Fashion MNIST data set [19]. The data $x$ enter the quantum circuit via a trainable encoding, i.e., they are encoded into unitary operators $U_{j,\Theta_j}(x)$ via an affine function $w_j^\top x + \theta_j$ with trainable parameters $w_j$, $\theta_j$. During training, we minimize a cost function consisting of the empirical loss as well as an additional regularization term penalizing the norms of the parameters $w_j$. This regularization reduces the Lipschitz bound of the quantum model with respect to data perturbations and, thereby, encourages improved robustness and generalization properties.

training schemes encouraging adversarial robustness of the resulting quantum model. Existing approaches in this direction include solving an (adversarial) min-max optimization problem during training [35] or adding adversarial examples to the training data set [43].

Besides robustness, another important aspect of any quantum model is its ability to generalize to unseen data [44,45]. In particular, various works have shown generalization bounds [24,46–49], i.e., bounds on the expected risk of a model depending on its performance on the training data. While these bounds provide insights into possibilities for constructing quantum models that generalize well, they also face inherent limitations due to their uniform nature [50].

### A. Contribution

This paper presents a flexible and rigorous framework for robustness and generalization of quantum models, providing both a theoretical analysis as well as a simple regularization strategy which allows to systematically adapt robustness and generalization during training (see Fig. 1 for an overview). More precisely, we first derive a Lipschitz bound of a given quantum model which explicitly involves the parameters of the data encoding. Based on this result, we propose a regularized training strategy penalizing the norm of the encoding parameters, which are considered trainable, in order to improve (adversarial) robustness of the model. Further, we derive a generalization bound which explicitly depends on the parameters of the quantum model and therefore does not share the limitations of existing *uniform* generalization bounds [50]. With numerical results, we demonstrate that the proposed Lipschitz bound regularization can indeed lead to substantial improvements in robustness and generalization of quantum models. Finally, given that the derived Lipschitz bound mainly depends on the norm of the data encoding, our results reveal the importance and benefits of trainable encodings over quantum circuits with a priori fixed encoding as frequently used in variational QML [20–24,28].

### B. Outline

The paper is structured as follows. In Sec. II, we introduce the considered class of quantum models with trainable encodings and we state their Lipschitz bound. Next, in Sec. III, we use the Lipschitz bound to study robustness of quantum models and to derive a regularization strategy for robust training whose benefits are demonstrated with numerical simulations. We then derive a generalization bound which depends explicitly on the data encoding parameters and we confirm this insight numerically by showing improved generalization under the proposed regularization strategy (Sec. IV). Further, in Sec. V, we discuss an important implication of our results on the benefits of trainable encodings for robustness and generalization. Finally, Sec. VI concludes the paper. In the Appendix, we provide technical proofs, details on the numerical simulations, as well as additional theoretical and numerical results.

## II. QUANTUM MODELS AND THEIR LIPSCHITZ BOUNDS

We consider parametrized unitary operators of the form

$$U_{j,\Theta_j}(x) = e^{-i(w_j^\top x + \theta_j)H_j}, \quad j =, 1 \ldots, N \tag{2}$$

with input data $x \in \mathbb{R}^d$, trainable parameters $\Theta_j = \{w_j, \theta_j\}$, $w_j \in \mathbb{R}^d$, $\theta_j \in \mathbb{R}$ and fixed Hermitian generators $H_j$. The generators $H_j$ are user-chosen, see Ref. [44] for references with guidelines. Depending on the choice of $H_j$, the operator $U_{j,\Theta_j}$ acts on either one or multiple qubits. The operators $U_{j,\Theta_j}$ give rise to the parametrized quantum circuit

$$U_\Theta(x) = U_{N,\Theta_N}(x) \cdots U_{1,\Theta_1}(x), \tag{3}$$

where $\Theta = \{\Theta_j\}_{j=1}^N$ comprises the set of trainable parameters. Throughout this paper, we abbreviate the $n_q$-qubit input state $|0\rangle^{\otimes n_q}$ by $|0\rangle$. The quantum model considered in this paper consists of $U_\Theta(x)$ applied to $|0\rangle$ and followed by a measurement with respect to the observable $\mathcal{M}$, i.e.,

$$f_\Theta(x) = \langle 0|U_\Theta(x)^\dagger \mathcal{M} U_\Theta(x)|0\rangle\,. \tag{4}$$

Note that each of the unitary operators $U_{j,\Theta_j}(x)$ involves the full data vector $x$, i.e., the data are loaded repeatedly into the circuit, a strategy that is commonly referred to as data reuploading [26]. The encoding of the data $x$ into each $U_{j,\Theta_j}(x)$ is realized via an affine function $w_j^\top x + \theta_j$, where both $w_j$ and $\theta_j$ are trainable parameters. Hence, we refer to (4) as a quantum model with trainable encoding. Such trainable encodings are a generalization of common quantum models [20–24,28], for which the $w_j$'s are fixed (typically unit vectors) and only the $\theta_j$'s are trained.

Our results rely on Lipschitz bounds (1). A Lipschitz bound quantifies the maximum perturbation of $f$ that can be caused by input variations. For the quantum model $f_\Theta$, we can state the following Lipschitz bound

$$L_\Theta = 2\|\mathcal{M}\| \sum_{j=1}^{N} \|w_j\|\|H_j\|. \tag{5}$$

The formal derivation can be found in Appendix A. For a given set of parameters $w_j$, (5) allows to compute the Lipschitz bound of the corresponding quantum model. Note that $L_\Theta$ depends only on $w_j$ but it is independent of $\theta_j$. This fact plays an important role for potential benefits of trainable encodings since the parameters $w_j$ are not optimized during training for fixed-encoding circuits. We note that all results in this paper hold for arbitrary $p$-norms as long as the same $p$ is used for both vector and induced matrix norms.

## III. ROBUSTNESS OF QUANTUM MODELS

Suppose we want to evaluate the quantum model $f_\Theta$ at $x$, i.e., we are interested in the value $f_\Theta(x)$, but we can only access $f_\Theta$ at some *perturbed* input $x' = x + \varepsilon$ with an unknown $\varepsilon$. Such a setup can arise due to various reasons, e.g., $x$ may be the output of some physical process which can only be accessed via noisy sensors. The perturbation $\varepsilon$ may also be the result of an adversarial attack, i.e., a perturbation aiming to cause a misclassification by choosing $\varepsilon$ such that

$$\|f_\Theta(x+\varepsilon) - f_\Theta(x)\| \tag{6}$$

is maximized. In either case, to correctly classify $x$ despite the perturbation, we require that $f_\Theta(x+\varepsilon)$ is close to $f_\Theta(x)$, meaning that (6) is small. According to (1), a Lipschitz bound $L$ of $f_\Theta$ quantifies exactly this difference, implying that the maximum possible deviation of $f_\Theta(x+\varepsilon)$ from $f_\Theta(x)$ is bounded as

$$\|f_\Theta(x+\varepsilon) - f_\Theta(x)\| \leqslant L\|\varepsilon\|. \tag{7}$$

This shows that smaller Lipschitz bounds imply better (worst-case) robustness of models against data perturbations. Thus, using (5), the robustness of the quantum model $f_\Theta$ is mainly influenced by the parameters of the data encoding $w_j$, $H_j$, and by the observable $\mathcal{M}$. In particular, smaller values of $\sum_{j=1}^{N} \|w_j\|\|H_j\|$ and $\|\mathcal{M}\|$ lead to a more robust model.

We now apply this theoretical insight to train robust quantum models using regularization. We consider a supervised learning setup with loss $\ell$ and training data set $(x_k, y_k) \in \mathcal{X} \times \mathcal{Y}$ of size $n$. The following optimization problem can be used to train the quantum model $f_\Theta$:

$$\min_\Theta \ \frac{1}{n}\sum_{k=1}^{n} \ell(f_\Theta(x_k), y_k). \tag{8}$$

In order to ensure that $f_\Theta$ not only admits a small training loss but is also robust and generalizes well, we add a regularization, leading to

$$\min_\Theta \ \frac{1}{n}\sum_{k=1}^{n} \ell(f_\Theta(x_k), y_k) + \lambda \sum_{j=1}^{N} \|w_j\|^2\|H_j\|^2. \tag{9}$$

Regularizing the parameters $w_j$ encourages small norms of the data encoding and, thereby, small values of the Lipschitz bound $L_\Theta$. We weight the parameter norms $\|w_j\|$ by $\|H_j\|$ due to their joint occurrence in (5). The hyperparameter $\lambda > 0$ allows for a trade off between the two objectives of a small training loss and robustness/generalization in the cost function. Note that the regularization does not involve the $\theta_j$'s since they do not influence the Lipschitz bound (5), an issue we discuss in more detail in Sec. V. Moreover, we do not introduce an explicit dependence of the regularization on $\mathcal{M}$ since we do not optimize over the observable in this paper. We note that penalty terms similar to the proposed regularization can be used for handling hard constraints in binary optimization via the quantum approximate optimization algorithm [51,52]. Indeed, the above regularization can be interpreted as a penalty-based relaxation of the corresponding constrained training problem, i.e., of training a quantum model with Lipschitz bound below a specific value.

We now evaluate our theoretical findings based on the circle classification problem from [53]: within a domain of $\mathcal{X} = [-1,+1] \times [-1,+1]$, a circle with radius $\sqrt{2/\pi}$ is drawn, and all data points inside the circle are labeled with $y = +1$, whereas points outside are labeled with $y = -1$, see Appendix C (Fig. 6). For the quantum model with trainable encoding, we use general SU(2) operators and encode $w_j^\top x + \theta_j$ into the first two rotation angles. We repeat this encoding for each of the considered 3 qubits, followed by nearest neighbor entangling gates based on CNOTs. Such a layer is then repeated 3 times. As observable, we use $\mathcal{M} = Z \otimes Z \otimes Z$. The resulting circuit is illustrated in Appendix C (Fig. 5). As norms in the regularized training problem (9), we employ 2-norms, but we note that exploring different choices is an interesting future direction. For example, regularization with nonsquared norms (e.g., a 1-norm) may enforce sparsity of the trained QML model and can, thereby, simplify its implementation on NISQ hardware.

The numerical results for the robustness simulations are shown in Fig. 2, where we compare the worst-case test accuracy and Lipschitz bound of three trained models with different regularization parameter $\lambda \in \{0, 0.2, 0.5\}$. Additionally, the plot shows the accuracy of a trained quantum model with fixed encoding for the same numbers of qubits and layers [see (14) and Appendix C for details]. The worst-case test accuracy of all models is obtained by sampling different noise samples $\varepsilon$ from $[-\bar{\varepsilon}, +\bar{\varepsilon}]$. This procedure amounts to finding adversarial noise samples and, therefore, the resulting worst-case test accuracy (approximately) quantifies the adversarial

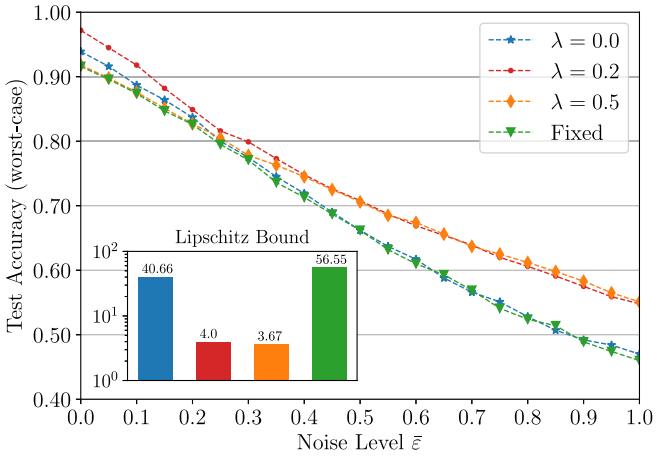


FIG. 2. We compare robustness of quantum models trained via (9) for $\lambda \in \{0, 0.2, 0.5\}$ and a quantum model with fixed encoding (14). As training and test set, we draw $n = 200$ and 1000 points $x_i \in \mathcal{X}$, respectively, uniformly at random. To study robustness, we perturb each of the 1000 test data points by random noise drawn uniformly from $[-\bar{\varepsilon}, +\bar{\varepsilon}]^d$ $(d = 2)$. The test accuracy in the plot is the worst case over 200 noise samples per data point.

robustness against attacks which are norm-bounded by $\bar{\varepsilon}$. As expected, all four models deteriorate with increasing noise level. For zero noise level $\bar{\varepsilon} = 0$, the model with the largest regularization parameter $\lambda = 0.5$ (and, hence, the smallest Lipschitz bound $L_\Theta = 3.67$) has a smaller test accuracy than the nonregularized model with $\lambda = 0$. This can be explained by a decrease in the training accuracy that is caused by the additional regularization in the cost. For increasing noise levels, however, the enhanced robustness outweighs the loss of training performance and, therefore, the model with $\lambda = 0.5$ outperforms the model with $\lambda = 0$. The fixed-encoding model achieves comparable performance to the trainable-encoding model with $\lambda = 0.5$ for small noise and the worst performance among all models for high noise. These observations can be explained by the high Lipschitz bound of the fixed-encoding model as well as its reduced expressivity, i.e., its limited ability to approximate functions from data due to the fixed encoding parameters $w_j$. Finally, the model with $\lambda = 0.2$ almost always outperforms the model with $\lambda = 0$ and, in particular, it yields a higher test accuracy for small noise levels. This can be explained by the improved generalization performance caused by the regularization, an effect we discuss in more detail in the following.

## IV. GENERALIZATION OF QUANTUM MODELS

The Lipschitz bound (5) not only influences robustness but also has a crucial impact on generalization properties of the quantum model $f_\Theta$. Intuitively, a smaller Lipschitz bound implies a smaller variability of $f_\Theta$ and, therefore, reduces the risk of overfitting. This intuition is made formal via the following generalization bound.

*Theorem IV 1* (Informal version). Consider a supervised learning setup with loss $\ell$ and data set $(x_k, y_k) \in \mathcal{X} \times \mathcal{Y}$ of size $n$ drawn according to the probability distribution $P$. For the quantum model $f_\Theta$ from (4), define the expected risk $\mathcal{R}(f_\Theta) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, f_\Theta(x)) dP(x, y)$ and the empirical risk $\mathcal{R}_n(f_\Theta) = \frac{1}{n}\sum_{k=1}^n \ell(y_k, f_\Theta(x_k))$. The generalization error of $f_\Theta$ is bounded as

$$|\mathcal{R}(f_\Theta) - \mathcal{R}_n(f_\Theta)| \leqslant C_1 \|\mathcal{M}\| \sum_{j=1}^N \|w_j\| \|H_j\| + \frac{C_2}{\sqrt{n}} \quad (10)$$

for some $C_1, C_2 > 0$.

The detailed version and proof of Theorem IV 1 are provided in Appendix B. Generalization bounds as in (10) quantify the ability of $f_\Theta$ to generalize beyond the available data. The bound (10) depends on the data encoding via $\sum_{j=1}^N \|w_j\| \|H_j\|$ and on the observable via $\|\mathcal{M}\|$. In particular, $f_\Theta$ achieves a small generalization error if its Lipschitz bound $L_\Theta$ is small and the size $n$ of the data set is large. Note, however, the following fundamental trade-off: A too small Lipschitz bound $L_\Theta$ may limit the expressivity of $f_\Theta$ and, therefore, lead to a high empirical risk $\mathcal{R}_n(f_\Theta)$, in which case the generalization bound (10) is meaningless. In conclusion, Theorem IV 1 implies a small expected risk $\mathcal{R}(f_\Theta)$ if $n$ is large, $L_\Theta$ is small, and $f_\Theta$ has a small empirical risk $\mathcal{R}_n(f_\Theta)$. In contrast to existing generalization bounds [24,46–49], the bound (10) is not uniform and explicitly involves the Lipschitz bound (5), i.e., the parameters of the data encoding. Hence, Theorem IV 1 does not share the limitations of uniform QML generalization bounds [50] and it can be used to systematically influence the generalization performance during training via regularization. In particular, according to Theorem IV 1, the regularized training problem (9) encourages models with improved generalization properties, where the hyperparameter $\lambda$ trades off between the empirical risk $\mathcal{R}_n(f_\Theta)$ and the generalization bound (10). Extending our results by studying the direct impact of $\lambda$ on the generalization performance is an interesting next step, which we expect to be nontrivial due to the nonconvexity of the loss function in (9), compare Ref. [54]. In practice, the hyperparameter $\lambda$ can be tuned, e.g., via cross-validation. We discuss and interpret the impact of the hyperparameter $\lambda$ in more detail with the following numerical results.

We evaluate the generalization performance of the trainable encoding again on the circle classification problem. The training setup is identical to the robustness simulations and the numerical results are shown in Fig. 3. Increasing the regularization parameter $\lambda$ decreases the Lipschitz bound $L_\Theta$ of the trained model. In accordance with the generalization bound (10), this reduction of $L_\Theta$ improves generalization performance with the maximum test accuracy at $\lambda = 0.15$. Beyond this value, the regularization causes a too small Lipschitz bound, limiting expressivity and, therefore, decreasing the training accuracy. As a result, the test accuracy decreases as well. This illustrates the role of $\lambda$ as a hyperparameter: Regularization does not always improve performance, but there is a sweet spot for $\lambda$ at which both superior generalization and robustness over the unregularized setup (i.e., $\lambda = 0$) can be obtained.

## V. BENEFITS OF TRAINABLE ENCODINGS

A popular class of quantum models is obtained by constructing circuits which alternate between data- and

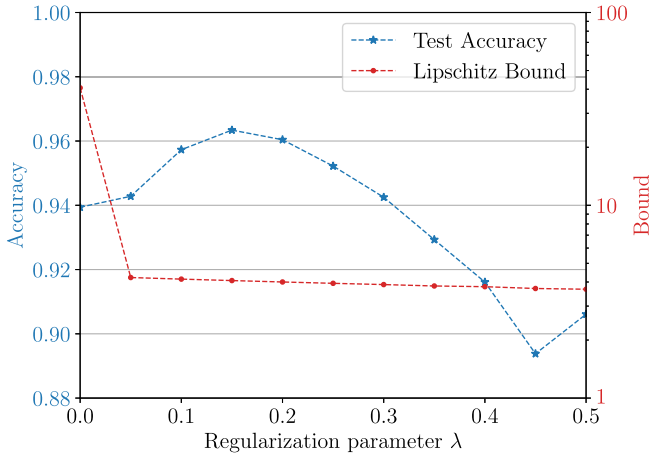


FIG. 3. Results for the generalization simulations. The training setup is identical to the robustness simulations as described in Fig. 2. As test set, we draw 10 000 points uniformly at random and evaluate the trained models with different regularization parameter λ.

parameter-dependent gates, i.e., replacing $U_\Theta(x)$ in (3) by

$$U_\phi^{\mathrm{f}}(x) = W(\phi_L)V(x)\cdots W(\phi_1)V(x), \tag{11}$$

compare [20–24,28]. The unitary operators $V$ and $W$ are given by

$$V(x) = e^{-ix_D G_D}\cdots e^{-ix_1 G_1}, \tag{12}$$

$$W(\phi_j) = e^{-i\phi_{j,p}S_p}\cdots e^{-i\phi_{j,1}S_1} \tag{13}$$

for trainable parameters $\phi_j$ and generators $G_i = G_i^\dagger$, $S_i = S_i^\dagger$. The corresponding quantum model is given by

$$f_\phi^{\mathrm{f}}(x) = \langle 0|U_\phi^{\mathrm{f}}(x)^\dagger \mathcal{M} U_\phi^{\mathrm{f}}(x)|0\rangle\,, \tag{14}$$

see Fig. 4.

It is not hard to show that the parametrized quantum circuit $U_\Theta(x)$ in (3) generalizes the one in (11). Indeed, $U_{j,\Theta_j}(x)$ in (2) reduces to either

$$e^{-ix_j G_j} \quad \text{or} \quad e^{-i\phi_j S_j} \tag{15}$$

for suitable choices of $w_j$, $\theta_j$, and $H_j$. Note that the data encoding of the quantum model $f_\phi^{\mathrm{f}}(x)$ is fixed a priori via the choice of $w_j$ and, in particular, it cannot be influenced during training. Therefore we refer to $f_\phi^{\mathrm{f}}(x)$ as a quantum model with fixed encoding, in contrast to $f_\Theta(x)$ in (4) which contains trainable parameters $w_j$ and, therefore, a trainable encoding.

Benefits of trainable encodings for the expressivity of quantum models have been demonstrated numerically in Refs. [55–57] and theoretically in Refs. [26,28]. In the following, we discuss the importance of trainable encodings for

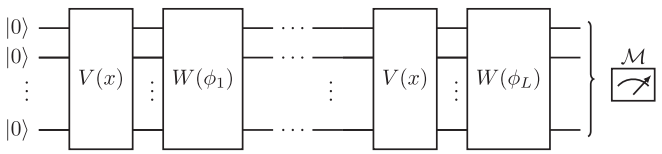


FIG. 4. Circuit representation of the quantum model (14) with fixed encoding.

robustness and generalization. Recall that the Lipschitz bound (5), which we showed to be a crucial quantifier of robustness and generalization of quantum models, only depends on the observable $\mathcal{M}$ and on the data encoding $w_j$, $H_j$, but is independent of the parameters $\theta_j$. Hence, in the quantum model $f_\phi^{\mathrm{f}}(x)$ with fixed encoding, the Lipschitz bound (5) cannot be influenced during training and, instead, is fixed a priori via the choice of the Hermitian generators $G_j$. As a result, training has a limited effect on robustness and generalization properties of fixed-encoding quantum models.

The distinction between trainable and fixed data encodings becomes even more apparent when expressing quantum models as Fourier series [28]. In this case, fixed-encoding quantum models choose the frequencies of the Fourier basis functions before the training and only optimize over their coefficients. On the contrary, trainable-encoding quantum models simultaneously optimize over the frequencies and the coefficients [57] which, according to the Lipschitz bound (5), is key for influencing robustness and generalization properties.

These insights confirm the observation by [58,59] that fixed-encoding quantum models are neither sensitive to data perturbations nor to overfitting. On the one hand, resilience against these two phenomena is a desirable property. However, the above discussion also implies that Lipschitz bound regularization, which is a systematic and effective tool for influencing robustness and generalization [2,6,16,18], cannot be implemented for fixed-encoding quantum models to improve robustness and generalization. Indeed, our robustness simulations in Fig. 2 show that the fixed-encoding model has a considerably higher Lipschitz bound than all the considered trainable-encoding models. This implies a significantly worse robustness with respect to data perturbations and, therefore, leads to a rapidly decreasing test accuracy for larger noise levels. Further, regularizing the parameters $\phi$ as suggested, e.g., by Ref. [60], does not affect the Lipschitz bound and, therefore, cannot be used to improve the robustness. In Appendix C, we study the effect of regularizing the $\phi_j$'s on generalization. We find that the influence of regularizing the $\phi_j$'s on the test accuracy is limited and likely dependent on the specific ground-truth distribution generating the data and the chosen circuit ansatz.

To conclude, our results show that training the encoding in quantum models not only increases the expressivity but also leads to superior robustness and generalization properties.

## VI. CONCLUSION

In this paper, we studied robustness and generalization properties of quantum models based on Lipschitz bounds. Lipschitz bounds are a well-established tool in the classical ML literature which not only quantify adversarial robustness but are also closely connected to generalization performance. We derived Lipschitz bounds based on the size of the data encoding which we then used to study robustness and generalization of quantum models. Given that our generalization bound explicitly involves the parameters of the data encoding, it does not face the limitations of uniform generalization bounds [50]. Further, our theoretical results highlight the role of trainable encodings combined with regularization techniques for obtaining robust and generalizable quantum models. The numerical results confirm our

theoretical findings, showing the existence of a sweet spot for the regularization parameter for which our training scheme improves both robustness as well as generalization compared to a nonregularized training scheme. It is important to emphasize that these numerical results with specific choices of rotation and entangling gates are mainly used for illustration, but our theoretical framework applies to all quantum models that can be written as (4) and, therefore, also allow for different rotation gates, entangling layers, or even parametrized multiqubit gates.

While our results indicate the potential of using Lipschitz bounds and regularization techniques in QML, it opens up various promising directions for future research. First and foremost, transferring existing research on Lipschitz bounds in classical ML to the QML setting provides a systematic framework for handling robustness and generalization, beyond the first results presented in this paper. For example, while we only consider quantum models with affine encodings $w_j^\top x + \theta_j$, it would be interesting to extend our results to more general, nonlinear encodings. Classical neural networks are ideal candidates for realizing a nonlinear encoding since their Lipschitz properties are well-studied [2,6,9–15], which would allow to train hybrid quantum-classical models which are not only expressive but also admit desirable robustness and generalization properties. Finally, although we focus on variational quantum models, the basic principles of our results are transferrable to different quantum models, including quantum kernel methods [22,23] or linear quantum models [25].

## ACKNOWLEDGMENTS

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-EXC 2075-390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). This work was also supported by the German Federal Ministry of Economic Affairs and Climate Action through the project AutoQML (Grant No. 01MQ22002A).

## DATA AVAILABILITY

The source code for the numerical case studies is publicly accessible on GitHub [27].

## APPENDIX A: LIPSCHITZ BOUNDS OF QUANTUM MODELS

In this section, we study Lipschitz bounds of quantum models as in (4). We first derive a Lipschitz bound which is less tight than the one in (5) but can be shown using a simple concatenation argument (Sec. A 1). Next, in Sec. A 2, we prove that (5) is indeed a Lipschitz bound.

### 1. Simple Lipschitz bound based on concatenation

Before stating the result, we introduce the notation

$$W = \begin{pmatrix} w_1^\top \\ \vdots \\ w_N^\top \end{pmatrix}, \quad \Omega = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_N \end{pmatrix}. \tag{A1}$$

*Theorem A 1.* The following is a Lipschitz bound of $f_\Theta$:

$$L = 2\|\mathcal{M}\|\|W\| \sum_{j=1}^{N} \|H_j\|. \tag{A2}$$

*Proof.* Our proof relies on the fact that a Lipschitz bound of a concatenated function can be obtained based on the product of the individual Lipschitz bounds. To be precise, suppose $f$ can be written as $f = f_1 \circ f_2 \circ \cdots \circ f_h$, where $\circ$ denotes concatenation and each $f_i$ admits a Lipschitz bound $L_i$, $i = 1, \ldots, h$. Then, for arbitrary input arguments $x, y$ of $f$, we obtain

$$\begin{aligned} \|f(x) - f(y)\| &\leqslant L_1 \|f_2 \circ \cdots \circ f_h(x) - f_2 \circ \cdots \circ f_h(y)\| \\ &\leqslant \cdots \leqslant L_1 L_2 \cdots L_h \|x - y\|. \end{aligned} \tag{A3}$$

We now prove that (A2) is a Lipschitz bound by representing $f_\Theta$ as a concatenation of the three functions

$$g_{\text{meas}}(z_{\text{m}}) = \langle z_{\text{m}}|\mathcal{M}|z_{\text{m}}\rangle\,, \tag{A4}$$

$$g_{\text{unitary}}(z_{\text{u}}) = e^{-iz_{\text{u},N}H_N} \cdots e^{-iz_{\text{u},1}H_1}\,|0\rangle\,, \tag{A5}$$

$$g_{\text{affine}}(z_{\text{a}}) = W z_{\text{a}} + \Omega. \tag{A6}$$

More precisely, it holds that

$$f_\Theta(x) = g_{\text{meas}} \circ g_{\text{unitary}} \circ g_{\text{affine}}(x). \tag{A7}$$

Hence, any set of Lipschitz bounds $L_{\text{meas}}, L_{\text{unitary}}, L_{\text{affine}}$ for the three functions $g_{\text{meas}}$, $g_{\text{unitary}}$, $g_{\text{affine}}$ gives rise to a Lipschitz bound of $f_\Theta$ as their product:

$$L = L_{\text{meas}} L_{\text{unitary}} L_{\text{affine}}. \tag{A8}$$

Therefore, in the following, we will derive individual Lipschitz bounds $L_{\text{meas}}$, $L_{\text{unitary}}$, and $L_{\text{affine}}$.

*Lipschitz bound of* $g_{\text{meas}}$. Note that

$$\frac{dg_{\text{meas}}(z_{\text{m}})}{dz_{\text{m}}} = 2\,\langle z_{\text{m}}|\,\mathcal{M}. \tag{A9}$$

Using $\|z_{\text{m}}\| = 1$, we infer

$$\left\| \frac{dg_{\text{meas}}(z_{\text{m}})}{dz_{\text{m}}} \right\| \leqslant 2\|\mathcal{M}\|. \tag{A10}$$

Thus $L_{\text{meas}} = 2\|\mathcal{M}\|$ is a Lipschitz bound of $g_{\text{meas}}$.

*Lipschitz bound of* $g_{\text{unitary}}$. It follows from [32, Theorem 2.2] that $L_{\text{unitary}} = \sum_{j=1}^{N} \|H_j\|$ is a Lipschitz bound of $g_{\text{unitary}}$.

*Lipschitz bound of* $g_{\text{affine}}$. Given the linear form of $g_{\text{affine}}$, we directly obtain that $L_{\text{affine}} = \|W\|$ is a Lipschitz bound.

### 2. Proof that (5) is a Lipschitz bound

We first derive a Lipschitz bound on the parametrized unitary $U_\Theta(x)$. To this end, we compute its differential

$$\begin{aligned} dU_\Theta(x) = (dU_{N,\Theta_N}(x))U_{N-1,\Theta_{N-1}}(x) \cdots U_{1,\Theta_1}(x) \\ + \cdots + U_{N,\Theta_N}(x) \cdots U_{2,\Theta_2}(x)(dU_{1,\Theta_1}(x)). \end{aligned} \tag{A11}$$

Note that each term $U_{j,\Theta_j}(x)$ can be written as the concatenation of the two maps $g_j$ and $h_j$ defined by

$$g_j(x) = w_j^\top x + \theta_j, \tag{A12}$$

$$h_j(z_j) = e^{-iz_j H_j}. \tag{A13}$$

To be precise, it holds that $U_{j,\Theta_j}(x) = h_j \circ g_j(x)$. The differentials of the two maps $g_j$ and $h_j$ are given by

$$dh_j(z_j)(u) = -iH_j e^{-iz_j H_j} u,$$
$$dg_j(x)(v) = w_j^\top v, \tag{A14}$$

where $h_j(z_j)(u)$ denotes the differential of $h_j$ at $z_j$ applied to $u \in \mathbb{R}$, and similarly for $g_j(x)(v)$. Thus we have

$$\begin{aligned} dU_{j,\Theta_j}(x)(v) &= (dh_j(g_j(x))) \circ (dg_j(x)(v)) \\ &= -iH_j e^{-i(w_j^\top x + \theta_j)H_j} w_j^\top v \\ &= -iH_j U_{j,\Theta_j}(x) w_j^\top v. \end{aligned} \tag{A15}$$

Inserting this into (A11), we obtain

$$\begin{aligned} &dU_\Theta(x)(v) \\ &\quad = -i\big(H_N U_\Theta(x) w_N^\top v + \cdots + U_\Theta(x) H_1 w_1^\top v\big). \end{aligned} \tag{A16}$$

We have thus shown that the Jacobian $J_\Theta(x)$ of the map $U_\Theta(x)\,|0\rangle$ is given by

$$J_\Theta(x) = -i\big(H_N U_\Theta(x)\,|0\rangle\, w_N^\top + \cdots + U_\Theta(x) H_1\,|0\rangle\, w_1^\top\big). \tag{A17}$$

Using that $|0\rangle$ has unit norm, that the $U_{j,\Theta_j}$'s are unitary, as well as the triangle inequality, the norm of $J_\Theta(x)$ is bounded as

$$\|J_\Theta(x)\| \leqslant \sum_{j=1}^{N} \|w_j\| \|H_j\|. \tag{A18}$$

Thus $\sum_{j=1}^{N} \|w_j\|\|H_j\|$ is a Lipschitz bound of $U_\Theta(x)\,|0\rangle$ [61, p. 356].

Finally, $f_\Theta(x)$ is a concatenation of $U_\Theta(x)\,|0\rangle$ and the function $z \mapsto \langle z|\mathcal{M}|z\rangle$, which admits the Lipschitz bound $2\|\mathcal{M}\|$, compare (A10). Hence, a Lipschitz bound of $f_\Theta$ can be obtained as the product of these two individual bounds, i.e., as in (5). ∎

## APPENDIX B: FULL VERSION AND PROOF OF THEOREM IV 1

We first state the main result in a general supervised learning setup, before applying it to the quantum model (4) considered in the paper. Consider a supervised learning setup with data samples $\{x_k, y_k\}_{k=1}^n$ drawn independently and identically distributed from $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ according to some probability distribution $P$. We define the $\epsilon$-covering number of $\mathcal{Z}$ as follows.

*Definition B 1* (adapted from [7, Definition 1]). We say that $\hat{\mathcal{Z}}$ is an *$\epsilon$-cover* of $\mathcal{Z}$, if, for all $z \in \mathcal{Z}$, there exists $\hat{z} \in \hat{\mathcal{Z}}$ such that $\|z - \hat{z}\| \leqslant \epsilon$. The *$\epsilon$-covering number* of $\mathcal{Z}$ is

$$\mathcal{N}(\epsilon, \mathcal{Z}) = \min\{|\hat{\mathcal{Z}}| \mid \hat{\mathcal{Z}} \text{ is an } \epsilon\text{-cover of } \mathcal{Z}\}. \tag{B1}$$

For a generic model $f : \mathcal{X} \to \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and the training data $\{x_k, y_k\}_{k=1}^n$, we define the expected loss and the empirical loss by

$$\mathcal{R}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, f(x)) dP(x, y) \tag{B2}$$

and

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{k=1}^{n} \ell(y_k, f(x_k)), \tag{B3}$$

respectively. The following result states a generalization bound of $f$.

*Lemma B 1.* Suppose

(1) the loss $\ell$ is nonnegative and admits a Lipschitz bound $L_\ell > 0$,

(2) $\mathcal{Z}$ is compact such that the value $M \coloneqq \sup_{y,y' \in \mathcal{Y}} \ell(y, y')$ is finite, and

(3) $L_f > 0$ is a Lipschitz bound of $f$.

Then, for any $\gamma, \delta > 0$, with probability at least $1 - \delta$ the generalization error of $f$ is bounded as

$$\begin{aligned} |\mathcal{R}(f) - \mathcal{R}_n(f)| &\leqslant \gamma L_\ell \max\{1, L_f\} \\ &\quad + M\sqrt{\frac{2\mathcal{N}\left(\frac{\gamma}{2}, \mathcal{Z}\right) \ln 2 + 2\ln\left(\frac{1}{\delta}\right)}{n}}. \end{aligned} \tag{B4}$$

*Proof.* For the following proof, we invoke the concept of $(K, \epsilon)$-robustness (adapted from [7, Definition 2]): The classifier $f$ is *$(K, \epsilon)$-robust* for $K \in \mathbb{N}$ and $\epsilon \in \mathbb{R}$, if $\mathcal{Z}$ can be partitioned into $K$ disjoint sets, denoted by $\{C_i\}_{i=1}^K$, such that the following holds: For all $k = 1, \ldots, n$, $(x, y) \in \mathcal{Z}$, $i = 1, \ldots, K$, if $(x_k, y_k), (x, y) \in C_i$, then

$$|\ell(y_k, f(x_k)) - \ell(y, f(x))| \leqslant \epsilon. \tag{B5}$$

This property quantifies robustness of $f$ in the following sense: The set $\mathcal{Z}$ can be partitioned into a number of subsets such that, if a newly drawn sample $(x, y)$ lies in the same subset as a testing sample $(x_k, y_k)$, then their associated loss values are close. Let us now proceed by noting that, for any $(x_k, y_k)$, $k = 1, \ldots, n$, and $(x, y) \in \mathcal{Z}$, it holds that

$$\begin{aligned} &|\ell(y_k, f(x_k)) - \ell(y, f(x))| \\ &\leqslant L_\ell \|(y_k, f(x_k)) - (y, f(x))\|_2 \\ &\leqslant L_\ell (\|y_k - y\| + \|f(x_k) - f(x)\|) \\ &\leqslant L_\ell \max\{1, L_f\} \|(x_k, y_k) - (x, y)\|, \end{aligned} \tag{B6}$$

where we use the Lipschitz bound $L_\ell$ of $\ell$, the triangle inequality, and the Lipschitz bound $L_f$ of $f$, respectively. Using [7, Theorem 6], we infer that $f$ is $(\mathcal{N}(\frac{\gamma}{2}, \mathcal{Z}), L_\ell \max\{1, L_f\}\gamma)$-robust for all $\gamma > 0$. It now follows from [7, Theorem 1] that, for any $\delta > 0$, with probability at least $1 - \delta$ inequality (B4) holds.

Let us now combine the Lipschitz bound (5) and lemma B 1 to state a tailored generalization bound for the considered class of quantum models $f_\Theta$, thus proving Theorem IV 1.

*Theorem B 1.* Suppose

(1) the loss $\ell$ is nonnegative and admits a Lipschitz bound $L_\ell > 0$ and

(2) $\mathcal{Z}$ is compact such that the value $M \coloneqq \sup_{y,y' \in \mathcal{Y}} \ell(y, y')$ is finite.

Then, for any $\gamma, \delta > 0$, with probability at least $1 - \delta$ the generalization error of $f_\Theta$ is bounded as

$$\begin{aligned} |\mathcal{R}(f) - \mathcal{R}_n(f)| &\leqslant \gamma L_\ell \max\left\{1, 2\|\mathcal{M}\| \sum_{j=1}^{N} \|w_j\| \|H_j\|\right\} \\ &\quad + M\sqrt{\frac{2\mathcal{N}\left(\frac{\gamma}{2}, \mathcal{Z}\right) \ln 2 + 2\ln\left(\frac{1}{\delta}\right)}{n}}. \end{aligned} \tag{B7}$$

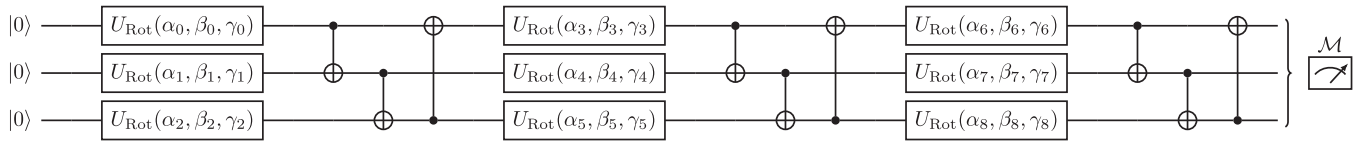


FIG. 5. The trainable encoding quantum model employed in our numerical case studies uses general $U_{\mathrm{Rot}} \in \mathrm{SU}(2)$ unitaries parametrized by three Euler angles $\alpha_i$, $\beta_i$, $\gamma_i$, that each have the form $w_i^\top x + \theta_i$ for trainable parameters $w_i$, $\theta_i$. We set $\gamma_j = 0$, $j = 0, \ldots, 8$, to enable a fair comparison to a quantum model with fixed encoding (since the data are two-dimensional, only two angles are needed for encoding the data via the latter).

Theorem B 1 shows that the size of the data encoding and of the observable directly influences the generalization performance of the quantum model $f_\Theta$. In particular, for smaller values of $\sum_{j=1}^N \|w_j\|\|H_j\|$ and $\|\mathcal{M}\|$, the expected loss is closer to the empirical loss. The right-hand side of (B7) contains two terms: The first one depends on the parameters of the quantum model $f_\Theta$ and characterizes its robustness via the derived Lipschitz bound, whereas the second term decays with increasing data length $n$. While (B4) holds for arbitrary values of $\gamma > 0$, it is not immediate which value of $\gamma$ leads to the smallest possible bound: smaller values of $\gamma$ decrease the first term but increase $\mathcal{N}(\frac{\gamma}{2}, \mathcal{Z})$ in the second term (and vice versa).

In contrast to existing QML generalization bounds [24,46–49], Theorem B 1 explicitly highlights the role of the model parameters via the Lipschitz bound. Using the additional flexibility of the parameter $\gamma$, it can be shown that the generalization bound (B7) converges to zero when the data length $n$ approaches infinity. To this end, we use Ref. [62, Lemma 6.27] to upper bound the covering number

$$\mathcal{N}\left(\frac{\gamma}{2}, \mathcal{Z}\right) \leqslant \left(\frac{6R}{\gamma}\right)^{d+1}, \tag{B8}$$

where $R$ is the radius of the smallest ball containing $\mathcal{Z}$. Inserting (B8) into (B7) and choosing $\gamma$ depending on $n$ as $\gamma = n^{-\frac{1}{2d+2}}$, we infer

$$|\mathcal{R}(f) - \mathcal{R}_n(f)| \leqslant \frac{1}{n^{\frac{1}{2d+2}}} L_\ell \max\left\{1, 2\|\mathcal{M}\| \sum_{j=1}^N \|w_j\|\|H_j\|\right\} + M\sqrt{\frac{2\ln 2(6R)^{d+1}}{\sqrt{n}} + \frac{2\ln\left(\frac{1}{\delta}\right)}{n}}, \tag{B9}$$

which indeed converges to zero for $n \to \infty$.

## APPENDIX C: NUMERICS: SETUP AND FURTHER RESULTS

In the following, we provide details regarding the setup of our numerical results (Sec. C 1) and we present further numerical results regarding parameter regularization in quantum models with fixed encoding (Sec. C 2).

### 1. Numerical setup

All numerical simulations within this work were performed using the PYTHON QML library PennyLane [63]. As device, we used the noiseless simulator "lightning.qubit" together with the adjoint differentiation method, to enable fast and memory efficient gradient computations. In order to solve the optimization problem in (9), we apply the ADAM optimizer using a learning rate of $\eta = 0.1$ and the suggested values for all other hyperparameters [64]. Furthermore, we run 200 epochs throughout and train 12 models based on different initial parameters for varying regularization parameters $\lambda \geqslant 0$. Adding the regularization does not introduce significant computational overhead as the evaluation of the cost only involves a weighted sum of the terms $\|w_j\|^2$. As final model, we take the set of parameters for the model with minimal cost over all runs and epochs. Furthermore, the training as well as the robustness and generalization analysis were parallelized using Dask [65].

For the trainable encoding model, the classical data is encoded into the quantum circuit with a general $U_{\mathrm{Rot}} \in \mathrm{SU}(2)$ unitary parametrized by 3 Euler angles $U_{\mathrm{Rot}}(\alpha_j, \beta_j, \gamma_j)$ with

$$\alpha_j = w_{j,1}^\top x + \theta_{j,1}, \tag{C1}$$

$$\beta_j = w_{j,2}^\top x + \theta_{j,2}, \tag{C2}$$

$$\gamma_j = w_{j,3}^\top x + \theta_{j,3}. \tag{C3}$$

$U_{\mathrm{Rot}}$ in PennyLane is implemented by the following decomposition:

$$U_{\mathrm{Rot}}(\alpha, \beta, \gamma) = R_Z(\gamma) R_Y(\beta) R_Z(\alpha), = \begin{pmatrix} e^{-\frac{i}{2}(\gamma+\alpha)} \cos\left(\frac{\beta}{2}\right) & -e^{\frac{i}{2}(\gamma-\alpha)} \sin\left(\frac{\beta}{2}\right) \\ e^{-\frac{i}{2}(\gamma-\alpha)} \sin\left(\frac{\beta}{2}\right) & e^{\frac{i}{2}(\gamma+\alpha)} \cos\left(\frac{\beta}{2}\right) \end{pmatrix}. \tag{C4}$$

In our numerical case study, we set $\gamma_j = 0$ for all $j$ since this (1) still allows to reach arbitrary points on the Bloch sphere and (2) enables an easier comparison to fixed-encoding quantum models. In order to introduce entanglement in a hardware-efficient way, we use a ring of CNOTs. The considered circuit is shown in Fig. 5 and involves three layers of rotations and entanglement, which we observed to be a good trade-off between expressivity and generalization. More precisely, in our simulations, fewer layers were not sufficient to accurately solve the classification task, whereas more layers led to higher degrees of overfitting. For the fixed-encoding quantum models as in (14), we use a similar three-layer ansatz, encoding the two entries of the 2D data points into the first and second angle $\alpha_j$, $\beta_j$ of the rotation gates, followed by a parametrized SU(2) rotation with three free parameters that are optimized, compare $\phi_j$ in (11). We also perform a classical data pre processing, scaling the input domain from $[-1, +1]$

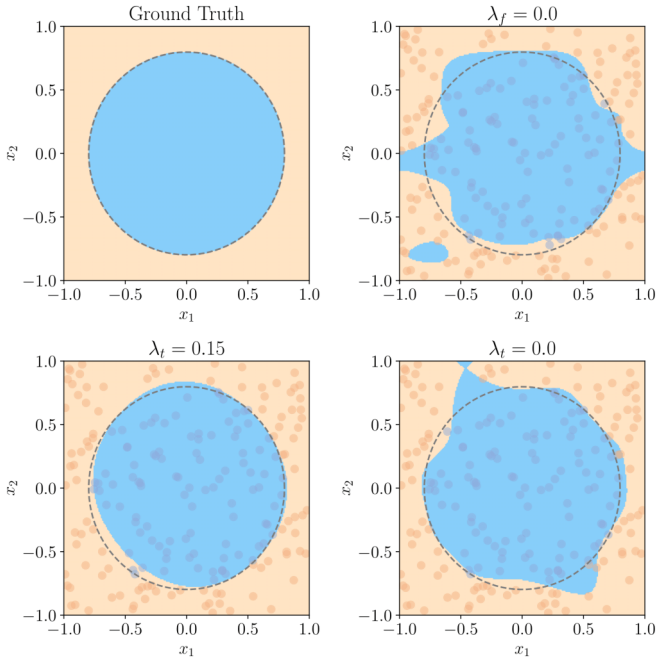


FIG. 6. From left to right, top to bottom: Illustration of the ground truth of the circle classification problem, the decision boundary for the fixed encoding model with regularization parameter $\lambda_f = 0.0$, the trainable encoding model with $\lambda_t = 0.15$ and with $\lambda_t = 0.0$. For the plot, we took the models with the lowest cost over all runs and epochs. Furthermore, the small circles denote the 200 training points.

to $[-\pi, +\pi]$, such that the full possible range of the rotation angles can be utilized.

In Fig. 6, we plot the ground truth and the decision boundaries for the two quantum models corresponding to $\lambda = 0.0$ and $\lambda = 0.15$, as well as the decision boundary for the fixed-encoding model. As expected, the decision boundary resulting from the regularized training is significantly smoother than the unregularized one, explaining the superior robustness and generalization of the former. Further, the fixed-encoding model does not accurately capture the ground truth due to its limited expressivity and high Lipschitz bound.

### 2. Regularization in quantum models with fixed encoding

In the main text, we have seen that the Lipschitz bound of the quantum model with fixed encoding $f^{\mathrm{f}}_{\phi}(x)$ in (14) cannot be adapted by changing the parameters $\phi$. As a result, it is not possible to use Lipschitz bound regularization for improving robustness and generalization. In the following, we investigate whether regularization of $\phi$ can instead be used to improve generalization performance. More precisely, we consider the same numerical setup as for our generalization results with

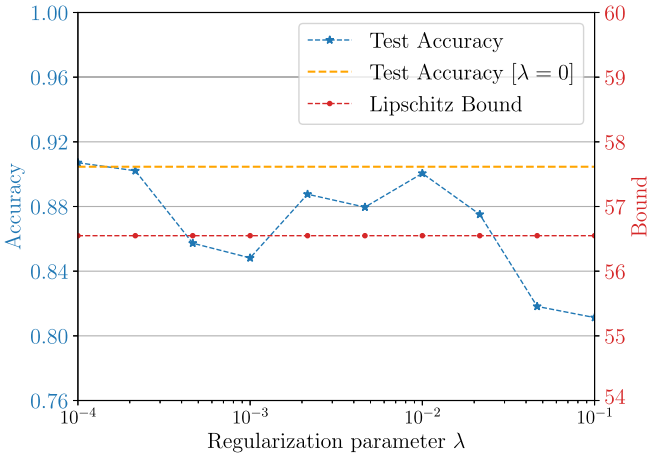


FIG. 7. Results for the generalization simulations for the circle classification problem when using the fixed-encoding quantum model (14) and regularization of the parameters $\phi$. The training and test setup are identical to simulations shown in Figs. 2 and 3. We plot the dependency of the test accuracy and the Lipschitz bound on the hyperparameter $\lambda$ entering the regularized training problem (C5). Further, we plot the test accuracy for the best solution of the unregularized training problem.

trainable encoding depicted in Fig. 3. The main difference is that we consider a fixed-encoding quantum model $f^{\mathrm{f}}_{\phi}(x)$ (compare Fig. 4) which is trained via the following regularized training problem

$$\min_{\phi} \frac{1}{n}\sum_{k=1}^{n} \ell\left(f^{\mathrm{f}}_{\phi}(x_k), y_k\right) + \lambda \sum_{j=1}^{L} \|\phi_j\|^2. \tag{C5}$$

The regularization with hyperparameter $\lambda > 0$ aims at keeping the norms of the angles $\phi_j$ small. Figure 7 depicts the test accuracy and Lipschitz bound of the resulting quantum model for different regularization parameters $\lambda$. First, note that the Lipschitz bound is indeed constant for all choices of $\lambda$ due to the fixed encoding. Comparing Figs. 3 and 7, we see that the trainable encoding yields a significantly higher test accuracy in comparison to the fixed encoding. Moreover, the influence of the regularization parameter on the test accuracy is much less pronounced for the fixed encoding than for the trainable encoding, confirming our previous discussion on the benefits of trainable encodings. The test accuracy is not entirely independent of $\lambda$ since (1) regularization of the parameters $\phi_j$ influences the optimization and can improve or deteriorate convergence and (2) biasing $\phi$ towards zero may be beneficial if the underlying ground-truth distribution is better approximated by a quantum model with small values $\phi$. Whether (2) brings practical benefits is, however, highly problem-specific as it depends on the distribution generating the data and on the circuit ansatz.

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, arXiv:1412.6572.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, arXiv:1312.6199.

[3] E. Wong and Z. Kolter, Provable defenses against adversarial examples via the convex outer adversarial polytope, in *Proceedings of the 35th International Conference on Machine Learning* (PMLR, 2018), pp. 5283–5292

[4] Y. Tsuzuku, I. Sato, and M. Sugiyama, Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks, in *Proceedings of the Advances in Neural Information Processing Systems* (PMLR, 2018), Vol. 80, pp. 6541–6550.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv:1706.06083.

[6] A. Krogh and J. Hertz, A simple weight decay can improve generalization, in *Advances in Neural Information Processing Systems* (PMLR, 1991), Vol. 4.

[7] H. Xu and S. Mannor, Robustness and generalization, Mach. Learn. **86**, 391 (2012).

[8] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in *Proceedings of the IEEE Symposium on Security and Privacy (SP)* (IEEE, Piscataway, NJ, 2016), pp. 582–597.

[9] U. von Luxburg and O. Bousquet, Distance-based classification with Lipschitz functions, J. Mach. Learn. Res. **5**, 669 (2004).

[10] P. Bartlett, D. J. Foster, and M. Telgarsky, Spectrally-normalized margin bounds for neural networks, in *Advances in Neural Information Processing Systems* (PMLR, 2017), Vol. 30, pp. 6240–6249.

[11] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, Exploring generalization in deep learning, in *Advances in Neural Information Processing Systems* (PMLR, 2017), pp. 5947–5956.

[12] J. Sokolić, R. Giryes, G. Sapiro, and M. R. D. Rodrigues, Robust large margin deep neural networks, IEEE Trans. Signal Process. **65**, 4265 (2017).

[13] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, Evaluating the robustness of neural networks: an extreme value theory approach, in *Proceedings of the 6th International Conference Learning Representations (ICLR)* (PMLR, 2018).

[14] W. Ruan, X. Huang, and M. Kwiatkowska, Reachability analysis of deep neural networks with provable guarantees, in *Proceedings of the 27th International Joint Conference Artificial Intelligence (IJCAI)* (AAAI Press, Stockholm, Sweden, 2018), pp. 2651–2659.

[15] C. Wei and T. Ma, Data-dependent sample complexity of deep neural networks via Lipschitz augmentation, in *Advances in Neural Information Processing Systems* (2019), pp. 9725–9736.

[16] M. Hein and M. Andriushchenko, Formal guarantees on the robustness of a classifier against adversarial manipulation, in *Proceedings of the Advances in Neural Information Processing Systems* (PMLR, 2017), pp. 2266–2276.

[17] H. Gouk, E. Frank, B. Pfahringer, and M. Cree, Regularisation of neural networks by enforcing Lipschitz continuity, Mach. Learn. **110**, 393 (2021).

[18] P. Pauli, A. Koch, J. Berberich, P. Kohler, and F. Allgöwer, Training robust neural networks using Lipschitz bounds, IEEE Control Syst. Lett. **6**, 121 (2022).

[19] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, arXiv:1708.07747.

[20] M. Schuld and F. Petruccione, *Machine Learning with Quantum Computers* (Springer, Berlin, 2021).

[21] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, Parameterized quantum circuits as machine learning models, Quantum Sci. Technol. **4**, 043001 (2019).

[22] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature (London) **567**, 209 (2019).

[23] M. Schuld and N. Killoran, Quantum machine learning in feature Hilbert spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[24] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks, Nat. Comput. Sci. **1**, 403 (2021).

[25] S. Jerbi, L. J. Fiderer, H. P. Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, Quantum machine learning beyond kernel methods, Nat. Commun. **14**, 517 (2023).

[26] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Data re-uploading for a universal quantum classifier, Quantum **4**, 226 (2020).

[27] https://github.com/daniel-fink-de/training-robust-and-generalizable-quantum-models.

[28] M. Schuld, R. Sweke, and J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, Phys. Rev. A **103**, 032430 (2021).

[29] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[30] R. LaRose and B. Coyle, Robust data encodings for quantum classifiers, Phys. Rev. A **102**, 032420 (2020).

[31] L. Cincio, K. Rudinger, M. Sarovar, and P. J. Coles, Machine learning of noise-resilient quantum circuits, PRX Quantum **2**, 010324 (2021).

[32] J. Berberich, D. Fink, and C. Holm, Robustness of quantum algorithms against coherent control errors, Phys. Rev. A **109**, 012417 (2024).

[33] D. Edwards and D. B. Rawat, Quantum adversarial machine learning: status, challenges and perspectives, in *Proceedings of the 2nd IEEE International Conference Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)* (IEEE, Piscataway, NJ, 2020), pp. 128–133.

[34] M. T. West, S.-L. Tsang, J. S. Low, C. D. Hill, C. Leckie, L. C. L. Hollenberg, S. M. Erfani, and M. Usman, Towards quantum enhanced adversarial robustness in machine learning, Nat. Mach. Intell. **5**, 581 (2023).

[35] S. Lu, L.-M. Duan, and D.-L. Deng, Quantum adversarial machine learning, Phys. Rev. Res. **2**, 033212 (2020).

[36] M. T. West, S. M. Erfani, C. Leckie, M. Sevior, L. C. L. Hollenberg, and M. Usman, Benchmarking adversarially robust quantum machine learning at scale, Phys. Rev. Res. **5**, 023186 (2023).

[37] N. Liu and P. Wittek, Vulnerability of quantum classification to adversarial perturbations, Phys. Rev. A **101**, 062331 (2020).

[38] H. Liao, I. Convy, W. J. Huggins, and K. B. Whaley, Robust in practice: adversarial attacks on quantum machine learning, Phys. Rev. A **103**, 042427 (2021).

[39] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, and N. Liu, Quantum noise protects quantum classifiers against adversaries, Phys. Rev. Res. **3**, 023153 (2021).

[40] J. Guan, W. Fang, and M. Ying, Robustness verification of quantum classifiers, in *Proceedings of the International Conference Computer Aided Verification* (Springer, Cham, 2021), pp. 151–174.

[41] M. Weber, N. Liu, B. Li, C. Zhang, and Z. Zhao, Optimal provable robustness of quantum classification via quantum hypothesis testing, npj Quantum. Inf. **7**, 76 (2021).

[42] W. Gong and D.-L. Deng, Universal adversarial examples and perturbations for quantum classifiers, Natl. Sci. Rev. **9**, nwab130 (2022).

[43] W. Ren, W. Li, S. Xu, K. Wang, W. Jiang, F. Jin, X. Zhu, J. Chen, Z. Song, P. Zhang, H. Dong, X. Zhang, J. Deng, Y. Gao, C. Zhang, Y. Wu, B. Zhang, Q. Guo, H. Li, Z. Wang, *et al.*, Experimental quantum adversarial learning with programmable superconducting qubits, Nat. Comput. Sci. **2**, 711 (2022).

[44] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, Challenges and opportunities in quantum machine learning, Nat. Comput. Sci. **2**, 567 (2022).

[45] E. Peters and M. Schuld, Generalization despite overfitting in quantum machine learning models, Quantum **7**, 1210 (2023).

[46] H.-Y. Huang, M. Bourghton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 2631 (2021).

[47] L. Banchi, J. Pereira, and S. Pirandola, Generalization in quantum machine learning: A quantum information standpoint, PRX Quantum **2**, 040321 (2021).

[48] M. C. Caro, E. Gil-Fuster, J. J. Meyer, J. Eisert, and R. Sweke, Encoding-dependent generalization bounds for parametrized quantum circuits, Quantum **5**, 582 (2021).

[49] S. Jerbi, C. Gyurik, S. C. Marshall, R. Molteni, and V. Dunjko, Shadows of quantum machine learning, Nat. Commun. **15**, 5676 (2024).

[50] E. Gil-Fuster, J. Eisert, and C. Bravo-Prieto, Understanding quantum machine learning also requires rethinking generalization, Nat. Commun. **15**, 2277 (2024).

[51] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[52] S. Hadfield, Z. Wang, E. G. Rieffel, B. O'Gorman, D. Venturelli, and R. Biswas, Quantum approximate optimization with hard and soft constraints, in *Proceedings of the Second International Workshop on Post Moores Era Supercomputing* (ACM Press, New York, NY, 2017), pp. 15–21.

[53] S. Ahmed, Tutorial: Data reuploading circuits (2021), https://pennylane.ai/qml/demos/tutorial_data_reuploading_classifier/.

[54] P. Huembeli and A. Dauphin, Characterizing the loss landscape of variational quantum algorithms, Quantum Sci. Technol. **6**, 025011 (2021).

[55] F. J. Gil Vidal and D. O. Theis, Input redundancy for parameterized quantum circuits, Front. Phys. **8**, 297 (2020).

[56] E. Ovalle-Magallanes, D. E. Alvarado-Carrillo, J. G. Avina-Cervantes, I. Cruz-Aceves, and J. Ruiz-Pinales, Quantum angle encoding with learnable rotation applied to quantum-classical convolutional neural networks, Appl. Soft Comput. **141**, 110307 (2023).

[57] B. Jaderberg, A. A. Gentile, Y. A. Berrada, E. Shishenina, and V. E. Elfving, Let quantum neural networks choose their own frequencies, Phys. Rev. A **109**, 042421 (2024).

[58] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[59] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).

[60] Y. Du, M.-H. Hsieh, T. Liu, S. You, and D. Tao, Learnability of quantum neural networks, PRX Quantum **2**, 040337 (2021).

[61] T. M. Apostol, *Mathematical Analysis*, 2nd ed. (Pearson Education, 1974).

[62] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning, Adaptive Computation and Machine Learning*, 2nd ed. (MIT Press, Cambridge, MA, 2018).

[63] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi, J. M. Arrazola, U. Azad, S. Banning, C. Blank, T. R. Bromley, B. A. Cordier, J. Ceroni, A. Delgado, O. D. Matteo, A. Dusko *et al.*, Pennylane: Automatic differentiation of hybrid quantum-classical computations, arXiv:1811.04968.

[64] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[65] Dask Development Team, Dask: Library for dynamic task scheduling, https://dask.org.