Classical surrogates for quantum learning models

Franz J. Schreiber,[1] Jens Eisert,[1,2,3] and Johannes Jakob Meyer[1]

[1]*Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, 14195 Berlin, Germany*
[2]*Helmholtz-Zentrum Berlin für Materialien und Energie, 14109 Berlin, Germany*
[3]*Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany*

(Dated: June 24, 2022)

The advent of noisy intermediate-scale quantum computers has put the search for possible applications to the forefront of quantum information science. One area where hopes for an advantage through near-term quantum computers are high is quantum machine learning, where variational quantum learning models based on parametrized quantum circuits are discussed. In this work, we introduce the concept of a *classical surrogate*, a classical model which can be efficiently obtained from a trained quantum learning model and reproduces its input-output relations. As inference can be performed classically, the existence of a classical surrogate greatly enhances the applicability of a quantum learning strategy. However, the classical surrogate also challenges possible advantages of quantum schemes. As it is possible to directly optimize the ansatz of the classical surrogate, they create a natural benchmark the quantum model has to outperform. We show that large classes of well-analyzed re-uploading models have a classical surrogate. We conducted numerical experiments and found that these quantum models show no advantage in performance or trainability in the problems we analyze. This leaves only generalization capability as possible point of quantum advantage and emphasizes the dire need for a better understanding of inductive biases of quantum learning models.

Quantum machine learning (QML) is a popular and widely studied application of quantum computers [1–3]. Theoretical evidence suggests that one day quantum machine learning methods can outperform classical computers in certain classical [4, 5] and quantum learning tasks [6]. Besides using quantum algorithms to train classical models [7–9], a particular emphasis is put on the construction of *quantum learning models*, which use quantum computers to parametrize hypothesis classes that are fit to the training data. Recently, much work has been done exploring *variational* models that use a *parametrized quantum circuit (PQC)* to make predictions, also referred to as *quantum neural networks (QNNs)*. While variational quantum models can be implemented on today's *noisy intermediate-scale quantum (NISQ)* devices [3, 10, 11], it is not clear if and how a practical quantum advantage can be realized within this framework [12].

An especially pressing issue when dealing with quantum learning models is the reliance on quantum hardware which severely limits how such models can be deployed in production environments. In this work, we argue that this challenge can be addressed if the quantum learning model in question has a *classical surrogate*, which we define as an equivalent classical model that can be obtained efficiently from a trained quantum learning model (see Fig. 1). The existence of a classical surrogate is a strong feature of a quantum learning model and can be considered as a fundamental prerequisite for any quantum learning model to be considered "practical". We show that a type of variational quantum re-uploading models [13] considered in Refs. [14–17] admits classical surrogates.

If a quantum model has a classical surrogate, quantum hardware is only required at the training stage. Any sort of quantum advantage therefore has to materialize at this stage through better training performance or generalization capability. In this setting the classical surrogate can also create a natural benchmark that the quantum model needs to beat to be relevant. This is the case when the classical surrogate itself can be turned into a learning model in its own right. We show that this is indeed the case for the re-uploading models considered in this work and numerically compare them with their classical surrogates on selected datasets.

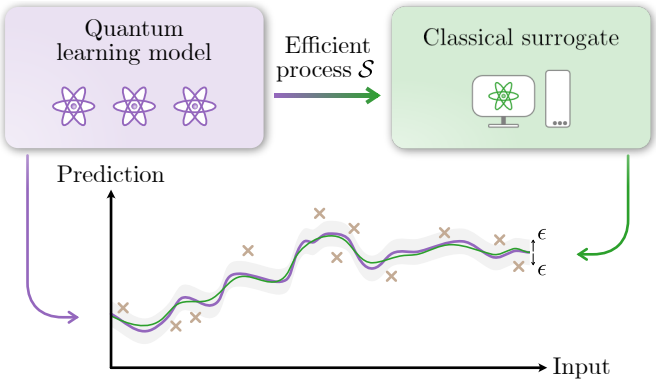

Figure 1. A quantum learning model has a classical surrogate if there exists a process that produces an equivalent classical model that is both efficient in the size of the quantum learning model and the desired approximation parameters.

In our experiments, we are unable to observe any advantage of the quantum learning model, neither in performance, training or generalization. We show that the classical surrogate can always achieve lower training loss and has a more favorable optimization landscape but we can not generally rule out the existence of an advantage in generalization capability. However, such an advantage would necessitate an understanding of the implementation of suitable inductive biases [18] in quantum learning models which is beyond the current state of the field.

Classical surrogates. Learning models executed on quan-

tum computers come with the inbuilt reliance on quantum hardware. This can impede their practicality tremendously – especially in the current NISQ-era where access to quantum computing resources is scarce. This situation is diametrically opposed to what makes classical machine learning attractive, where training the model might be very demanding, but obtaining new predictions is simple and can be done on less powerful client-side devices.

We can circumvent this impediment if we have access to a classical model that reproduces the same input-output relations as the quantum learning model. We call such a classical replacement of the trained quantum learning model a *classical surrogate*. A essential prerequisite for such a definition to be non-trivial is *efficiency* in the process that creates the surrogate from the trained quantum learning model as well as in the evaluation and storage of the classical surrogate itself. We can not expect that classical surrogates generically exist for all quantum learning models. Due to the concerns outlined above, having a classical surrogate is thus a fundamental property of any quantum learning model that can be considered "practical".

Formally, we define a classical surrogate for a hypothesis class $\mathcal{F}$ of quantum learning models with inputs $\boldsymbol{x} \in \mathcal{X}$ and outputs $y \in \mathcal{Y}$ as follows:

Definition 1 (Classical surrogate)**.** *A hypothesis class of quantum learning models $\mathcal{F}$ has* classical surrogates *if there exists a process $\mathcal{S}$ that upon input of a learning model $f \in \mathcal{F}$ produces a classical model $g_f \in \mathcal{G}$ such that the maximal deviation of the surrogate from the quantum learning model is bounded with high probability. Formally, we require*

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{X}} \|f(\boldsymbol{x}) - g_f(\boldsymbol{x})\| \leq \epsilon\right] \geq 1-\delta, \tag{1}$$

for a suitable norm on the output space $\mathcal{Y}$. The surrogation process $\mathcal{S}$ must me efficient in the size of the quantum learning model, the error bound ϵ and the failure probability δ.

We consider the supremum norm for the deviation to be a necessary feature, as more coarse-grained notions of approximation could tolerate isolated "outliers" for which the surrogate produces very different outputs than the quantum model.

Interestingly, the existence of a classical surrogate immediately implies that any advantage of the quantum learning model must be realized at the training stage, either through significant speedups, increases in training performance, cost reductions or better generalization capabilities. For such advantages, the classical surrogate can provide a natural benchmark if it can be turned into a learning model in its own right. In this way, quantum learning models with classical surrogates can be amenable to a notion of "dequantization".

Variational re-uploading models. We consider the same type of variational quantum learning model as Refs. [14–17] where vector-valued inputs $\mathcal{X} = \mathbb{R}^d$ are mapped to real outputs $\mathcal{Y} = \mathbb{R}$. The model is constructed by applying L layers of trainable unitaries $W^{(j)}(\boldsymbol{\theta})$ interleaved with data-encoding blocks $S^{(j)}(\boldsymbol{x})$ resulting in a parametrized circuit

$$U(\boldsymbol{x},\boldsymbol{\theta}) = W^{(L)}(\boldsymbol{\theta})S^{(L)}(\boldsymbol{x})\ldots W^{(1)}(\boldsymbol{\theta})S^{(1)}(\boldsymbol{x})W^{(0)}. \tag{2}$$

Predictions are obtained by evaluating the expectation value of an arbitrary observable M with bounded operator norm after $U(\boldsymbol{x},\boldsymbol{\theta})$ is applied to the all-zero state:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle 0|U^\dagger(\boldsymbol{x},\boldsymbol{\theta})MU(\boldsymbol{x},\boldsymbol{\theta})|0\rangle. \tag{3}$$

In Refs. [15, 16], it has been shown that learning models of this type can be expanded into a truncated Fourier series

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}\in\Omega} c_{\boldsymbol{\omega}}(\boldsymbol{\theta})e^{-i\boldsymbol{\omega}\boldsymbol{x}}, \tag{4}$$

where the set of accessible frequencies Ω depends only on the structure of the $S^{(j)}(\boldsymbol{x})$ and the number of layers.

In the following, we will assume that the data encodings $S^{(j)}(\boldsymbol{x})$ are composed of elementary data encodings of the form $S_k^{(j)}(x_k) = \exp(-ix_kH_k^{(j)})$ with $H_k^{(j)}$ having integer eigenvalue differences, which means that all data features are elementary parameters of rotation gates [19]. The integer eigenvalue differences guarantee that the model output is periodic on the interval $[0, 2\pi)$. It has been shown in Ref. [17] that for models of this type with constrained locality of the gates $S_k^{(j)}(x_k)$, the number of accessible frequencies as well as the maximal frequency grows only polynomially in the number of encoding gates N and hence also at most polynomial in the number of qubits.

Fourier-based classical surrogates. We can explicitly exploit the fact that the outputs of the model are guaranteed to be truncated Fourier series with known frequencies to construct a classical surrogate for these models. We denote the Fourier-based surrogate as

$$g_{\boldsymbol{c}}(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}\in\Omega} c_{\boldsymbol{\omega}}e^{-i\boldsymbol{\omega}\boldsymbol{x}} \tag{5}$$

where the Fourier coefficients $\boldsymbol{c} = (c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega}$ are the parameters that need to be computed with a guarantee that fulfills the surrogation conditions of Definition 1.

To do so, we use the following protocol based on the discrete Fourier transform: For each of the d data features, set $T_i = 2\omega_{\max}(i) + 1$ where $\omega_{\max}(i) = \max\{|\omega_i| \colon \boldsymbol{\omega} \in \Omega\}$ is the maximal frequency for the i-th data feature. We use this to define an equally-spaced grid on the interval $[0, 2\pi)$ with T_i points for every data feature which yields a grid for the whole set with $T = \prod_{i=1}^d T_i$ elements. For every datapoint $\boldsymbol{x}_j$ in this grid, we obtain N samples from the quantum model's output and compute an estimate for the expectation value through the corresponding sample mean $\hat{y}_j$. We then solve the least-squares problem

$$\boldsymbol{c}_* = \operatorname*{argmin}_{\boldsymbol{c}} \sum_{j=1}^{T} \left|g_{\boldsymbol{c}}(\boldsymbol{x}_j) - \hat{y}_j\right|^2. \tag{6}$$

We transform this into a linear system by defining

$$A_{j,\boldsymbol{\omega}} = e^{-i\boldsymbol{\omega}\boldsymbol{x}_j},\ \hat{\boldsymbol{y}}_j = y_j, \tag{7}$$

with which it reduces to

$$\boldsymbol{c}_* = \underset{\boldsymbol{c}}{\operatorname{argmin}} \|A\boldsymbol{c} - \hat{\boldsymbol{y}}\|^2. \tag{8}$$

The solution for this problem can be computed via the pseudo-inverse using a singular value decomposition or as a convex program. We have the following recovery guarantee for the whole protocol:

Proposition 1. *The classical surrogate $g_{\boldsymbol{c}_*}$ obtained through the above protocol fulfills the surrogation conditions of Definition 1 using a total of*

$$N_{\text{total}} = TN = \frac{2T\|M\|_\infty^2}{\epsilon^2}\left(\log\frac{1}{\delta} + T\log 2\right) \tag{9}$$

invocations of the quantum learning model.

Proof. The full proof is relegated to Appendix A for brevity. It leverages transportation-cost inequalities to show concentration of the ℓ_1-norm approximation of the Fourier coefficients. The result follows from combining this with known guarantees for the discrete Fourier transform. □

The surrogation protocol only has a sub-linear overhead in T compared to the sample complexity of conducting inference with the quantum model as obtaining the output for T different inputs to accuracy ϵ with probability at least $1-\delta$ has a sample complexity of

$$N_{\text{inference}} = \frac{2T\|M\|_\infty^2}{\epsilon^2}\log\frac{2T}{\delta} \tag{10}$$

We provide a proof of this well-known fact in Appendix B for completeness.

As $T = O(\omega_{\max}^{d/2})$, which is polynomial in the number of qubits, the above protocol is efficient in all relevant variables, as required by the surrogation conditions. While the number of data features d is constant for a given learning problem, the exponential scaling in this variable can present a challenge to scale up the classical surrogate. It is an intriguing question how further structural assumptions could be used to improve upon the above protocol.

We now outline how we can directly train the classical surrogates as learning models. Nothing prevents us from directly minimizing the least-squares loss of Eq. (6) for the given training data. We could use the same strategy as for the construction of the classical surrogate and solve the linear system through a singular value decomposition. For larger problems this, however, becomes impractical in time and memory requirements and the matrix A could be ill-conditioned as we discuss in Appendix C. Furthermore, a perfect solution to the problem is usually not desirable to avoid overfitting.

We therefore opt to use stochastic gradient descent methods on the Fourier coefficients, which avoids the aforementioned problems. The convexity of the optimization problem furthermore guarantees convergence to the global optimum if the learning rate is suitably parametrized [20]. We implement this model as a neural network which facilitates backpropagation through the surrogate model, which means we can apply the same optimization techniques to quantum model and classical surrogate. The unrestricted optimization of the Fourier coefficient implies that the global optimum of the classical surrogate provides a lower bound for the global training loss achievable by the quantum model, but also means that there is a higher danger of overfitting which we mitigate by observing the validation loss. Additionally, the loss landscapes of the quantum model and the classical surrogate can differ dramatically because the linear least-squares problem is convex, whereas the loss landscapes of quantum learning models are usually rugged and complicated [21], especially at low parameter counts before overparametrization phenomena kick in [22, 23].

Numerical implementation. In this section, we compare re-uploading models based on parametrized quantum circuits to their classical surrogates. We emphasize that there are countless possibilities to tweak the performance of the classical surrogate, like regularization or reparametrization. The point of this section, however, is to find the simplest model that matches or beats the corresponding quantum learning model.

The quantum model we consider is a re-uploading model where each layer of the model consists of a data encoding block $S(\boldsymbol{x}) = \bigotimes_{i=1}^{d} R_X(x_i)$ and a trainable block $W^{(l)}$, $L \in \{0, \ldots, L\}$. For the trainable block $W^{(l)}$ we choose the *Strongly Entangling Layer* template provided by *PennyLane* which consists of B block layers, where B – as well as the number of total layers L – is a hyperparameter of the quantum learning model (see Appendix D 1). In the following we use models with $B \in \{1, 3\}$ and $L \in \{2, 3\}$. The surrogate model is implemented as a neural network with one linear layer with the Fourier coefficients as weights. To avoid dealing with complex numbers, we have used the equivalent expansion in terms of cosines and sines. Note that the set of accessible frequencies only depends on L and that increasing B only increases the expressivity of the Fourier coefficients.

In our implementation, we use *PennyLane* [24] for the quantum parts and *PyTorch* [25] for the classical parts. The problems considered here are all of moderate size, therefore we use the memory-intensive LBFGS algorithm which is guaranteed to converge under the Wolf conditions, which spares the search for suitable learning rates [26].

Results. As we have already outlined, advantages of a quantum learning model could come in many flavors, be it easier trainability, better generalization or significant speedups. We compare the performance of quantum learning models with two different numbers of parameters and the corresponding classical surrogate on three different learning problems: A synthetic dataset generated using the *make_regression* function provided by *Scikit-learn* [27], a dataset obtained from sampling outputs of a randomized quantum re-uploading model and the California housing dataset which is a standard benchmark dataset for regression tasks. We present the loss

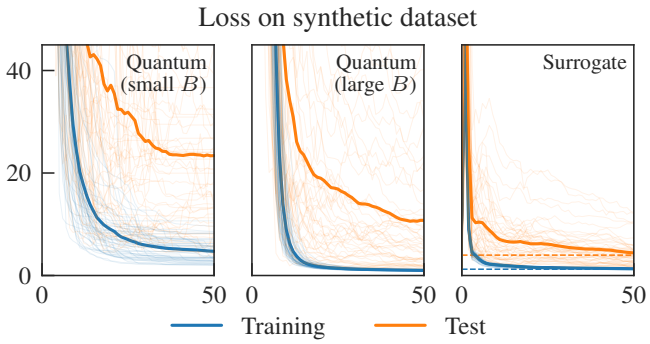

Figure 2. Training and test loss over epochs for three different learning models trained on a standard synthetic dataset over random splits. The dashed lines indicate the mean loss associated with the empirical risk minimizer. We observe the effects of higher expressivity when the number of parameters of the quantum model is increased but the surrogate model outperforms both quantum models in loss and trainability.

curves for the synthetic dataset in Fig. 2, the loss curves for the other problems can be found in Appendix D. As we used a small dataset for the synthetic case, we randomized the train-test-split for each run to avoid results that depend on a particular split.

As expected, we observe that the classical surrogate consistently achieves lower training loss as the quantum models and that the training loss converges as nicely as one would expect from a convex problem. We further witness that an increase in the number of block layers B, and hence a greater expressivity, allows the quantum model to perform better, both in test and training loss as well as in trainability in the synthetic and random PQC learning tasks. The increased performance in trainability is observed by the lower variance and increased smoothness of training curves. This indicates that the loss landscapes at low expressivity are highly frustrated which is in good accordance with recent results on loss landscape of variational models [21–23, 28].

This observation, however, also points to the fact that with increased expressivity, the expected behavior of the quantum models also tends to be more and more similar to the one of the classical surrogate, which – in a way – presents an idealized limit of the quantum model. However, while this is desirable from the perspective of training performance, it also increases the danger that the quantum model looses its ability to encode an inductive bias that is different from the one of the classical surrogate in a meaningful way, in which case the direct optimization of the surrogate is usually the better alternative. This immediately raises the question if there even exists a "sweet spot" where quantum re-uploading models of the type considered in this work are advantageous, as they interpolate between a setting of high bias but low trainability and a setting of low bias and high trainability.

The fact that we do not observe any kind of advantage of the quantum learning model over the classical surrogate in the examples we study can well be a limitation of the particular parametrizations of the quantum learning models that we consider, which are built on circuit templates available in the literature. We can therefore not make a statement about the ultimate capabilities of these models, as for this it is first necessary to better understand the relation of circuit structures and the corresponding inductive biases of the quantum learning model. However, we can conclude that for quantum learning models constructed from contemporary circuit ansätze, engineering an equivalent or better classical model is rather simple.

Conclusion. In this work, we have introduced the concept of a *classical surrogate* for a quantum learning model. Having access to a classical replacement that can be efficiently constructed from a trained quantum learning model greatly enhances its applicability and interpretability. We have shown that a widely analyzed type of re-uploading models has a classical surrogate. This is possible because this class of re-uploading models can be expanded in terms of a truncated Fourier series with a modest number of coefficients, the classical surrogate is then also a truncated Fourier series whose coefficients can be found efficiently by performing a discrete Fourier transform.

Classical surrogates have utility beyond removing the need of a quantum device for their use in production environments. They provide a natural benchmark by offering a concrete and natural test for any claim of "quantum advantage": A quantum learning model can not exhibit a quantum advantage if it does not possess trainability, expressivity or generalization properties superior to its classical surrogate. Therefore, classical surrogates can be used as a tool to pin down regimes where a possible quantum advantage could occur by indicating when the quantum model enters a "classical regime" where one could just equivalently train the classical surrogate. Conversely, impracticality of the classical surrogate can indicate a possible regime of advantage.

Applying these concepts to selected simple learning problems, we have observed that re-uploading models constructed from contemporary circuit ansätze can be beat by a simple classical surrogate model as we did not witness advantages in training, performance or an inductive bias towards favorable solutions. It is still conceivable that such an inductive bias could exist, but our understanding of its relation to particular circuit templates is too ill-understood to realize it. It is our hope that this work stimulates further research into the precise potential of variational quantum circuit for learning tasks and into classical surrogates for other classes of quantum learning models.

Acknowledgments. We would like to thank Sofiene Jerbi, Ingo Roth and Daniel Stilck-França for insightful discussions. We thank the BMBF (Hybrid), the BMWK (PlanQK), the QuantERA (HQCC), the Munich Quantum Valley (K8) and the Einstein Foundation (Einstein Research Unit on Quantum Devices) for their support.

[1] M. Schuld, I. Sinayskiy, and F. Petruccione, Contemp. Phys. **56**, 172 (2015).

[2] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Nature **549**, 195 (2017).

[3] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Rev. Mod. Phys. **94**, 015004 (2022).

[4] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, Quantum **5**, 417 (2021).

[5] Y. Liu, S. Arunachalam, and K. Temme, Nat. Phys. **17**, 1013 (2021).

[6] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, and J. R. McClean, arXiv:2112.00778 (2021).

[7] P. Rebentrost, M. Mohseni, and S. Lloyd, Phys. Rev. Lett. **113**, 130503 (2014).

[8] G. Verdon, J. Pye, and M. Broughton, arXiv:1806.09729 (2018).

[9] T. Hubregtsen, C. Segler, J. Pichlmeier, A. Sarkar, T. Gabor, and K. Bertels, in *2020 21st International Symposium on Quality Electronic Design (ISQED)* (2020) pp. 329–334, ISSN: 1948-3287.

[10] T. Jones, S. Endo, S. McArdle, X. Yuan, and S. C. Benjamin, Phys. Rev. A **99**, 062304 (2019).

[11] M. Lubasch, J. Joo, P. Moinier, M. Kiffner, and D. Jaksch, Phys. Rev. A **101**, 010301 (2020).

[12] M. Schuld and N. Killoran, arXiv:2203.01340 (2022).

[13] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Quantum **4**, 226 (2020).

[14] J. G. Vidal and D. O. Theis, arXiv:1812.06323 (2018).

[15] F. J. G. Vidal and D. O. Theis, Front. Phys. **8** (2020), 10.3389/fphy.2020.00297.

[16] M. Schuld, R. Sweke, and J. J. Meyer, Phys. Rev. A **103**, 032430 (2021).

[17] M. C. Caro, E. Gil-Fuster, J. J. Meyer, J. Eisert, and R. Sweke, Quantum **5**, 582 (2021).

[18] J. M. Kübler, S. Buchholz, and B. Schölkopf, arXiv:2106.03747 (2021).

[19] Through this we isolate the properties of the underlying quantum model in our analysis. If pre-processing needs to be considered, then one could always use the different functions of the input variables as new parameters.

[20] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade: Second edition*, edited by G. Montavon, G. B. Orr, and K.-R. Müller (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012) pp. 421–436.

[21] M. S. Rudolph, S. Sim, A. Raza, M. Stechly, J. R. McClean, E. R. Anschuetz, L. Serrano, and A. Perdomo-Ortiz, arXiv:2111.04695 (2021).

[22] M. Larocca, N. Ju, D. García-Martín, P. J. Coles, and M. Cerezo, arXiv:2109.11676 (2021).

[23] E. R. Anschuetz, arXiv:2109.06957 (2022).

[24] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, M. S. Alam, S. Ahmed, J. M. Arrazola, C. Blank, A. Delgado, S. Jahangiri, K. McKiernan, J. J. Meyer, Z. Niu, A. Száva, and N. Killoran, arXiv:1811.04968 (2018).

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.

[26] D. C. Liu and J. Nocedal, Math. Prog. **45**, 503 (1989).

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[28] R. Wiersema, C. Zhou, Y. de Sereville, J. F. Carrasquilla, Y. B. Kim, and H. Yuen, PRX Quantum **1**, 020319 (2020).

[29] M. Raginsky and I. Sason, arXiv:1212.4663 (2021).

[30] A. L. Gibbs and F. E. Su, arXiv:math/0209021 (2020).

[31] D. Nagel, "The condition number of Vandermonde matrices and its application to the stability analysis of a subspace method," (2020), PhD thesis, Universität Osnabrück.

[32] M. Kliesch and I. Roth, PRX Quantum **2**, 010201 (2021).

Appendix A: Recovery guarantees

In our reconstruction of the Fourier coefficients, we have to take into account that we have finite sampling statistics when we compute expectation values from the outputs of the quantum device. To do so, we first start by proving a concentration bound for a vector of sample means.

Lemma 1. *Let $\hat{\boldsymbol{\xi}} = (\hat{\xi}_{i,j})_{i=1\,j=1}^{T\quad N} \in \mathcal{X}$ be a collection of i.i.d. zero-centered random variables such that $|\hat{\xi}_{i,j}| \leq B$ for all i and j. Let $\hat{\boldsymbol{\eta}}$ denote the vector of the T sample means $\hat{\eta}_i = \frac{1}{N}\sum_{j=1}^{N} \hat{\xi}_{i,j}$. The ℓ_1-norm of $\hat{\boldsymbol{\eta}}$ then obeys the large deviation bound*

$$\mathbb{P}\left[\|\hat{\boldsymbol{\eta}}\|_1 \geq \alpha\right] \leq \exp\left(\log(2)T - \frac{\alpha^2 N}{2TB^2}\right). \tag{A1}$$

Before we proceed to the proof of the above Lemma, we note that this is an improvement over what could be obtained by element-wise application of Hoeffding's inequality which would yield a right hand side of

$$\exp\left(\log(2)\log(2T) - \frac{\alpha^2 N}{2T^2B^2}\right). \tag{A2}$$

The proof of Lemma 1 relies on Gaussian concentration inequalities obtained from transportation cost inequalities. To state the main theorem we build on, we first need a definition:

Definition 2. *A probability measure μ on a space $\mathcal{X}$ with distance measure d satisfies $T_1(c)$ if the 1-Wasserstein distance of μ to any other measure ν on $\mathcal{X}$,*

$$W_1(\mu,\nu) = \inf\left\{\int_{\mathcal{X}\times\mathcal{X}} \mathrm{d}\pi(x,y)\, d(x,y) \;\middle|\; \int_{\mathcal{X}} \mathrm{d}y\, \pi(x,y) = \mu(x), \int_{\mathcal{X}} \mathrm{d}x\, \pi(x,y) = \nu(y)\right\}, \tag{A3}$$

obeys an upper bound through the relative entropy of the form

$$W_1(\mu,\nu) \leq \sqrt{2cD(\nu\|\mu)}. \tag{A4}$$

A measure that fulfills $T_1(c)$ has nice concentration properties as is witnessed by the following Theorem:

Theorem 1 (Corollary 3.4.1 of Ref. [29])**.** *Let μ be a probability distribution over $\boldsymbol{\xi} \in \mathcal{X}$ that satisfies $T_1(c)$ and F be a Lipschitz function with Lipschitz constant $\|F\|_{\mathrm{Lip}}$. Then*

$$\mathbb{P}\left[F(\boldsymbol{\xi}) - \mathbb{E}[F(\boldsymbol{\xi})] \geq \alpha\right] \leq \exp\left(-\frac{\alpha^2}{2c\|F\|_{\mathrm{Lip}}^2}\right). \tag{A5}$$

Before we can come to the final proof, we establish that the type of measure we are interested in fulfills $T_1(c)$ due to the bounded nature of the involved variables:

Lemma 2. *Let $\{\mu_i\}_{i=1}^n$ be measures on the interval $\mathcal{I} = [-B, B]$ equipped with the distance measure $d(x,y) = |x-y|$. Then, $\bigotimes_{i=1}^n \mu_i$ as a measure over $\mathcal{I}^n$ is $T_1(nB^2)$ with respect to $d(x_n, y_n) = \|x_n - y_n\|_1$.*

Proof. We can upper-bound the 1-Wasserstein distance for μ to any other measure on $\mathcal{I}$ through the diameter of $\mathcal{I}$ [30]

$$W_1(\mu,\nu) \leq \mathrm{diam}(\mathcal{I}) d_{\mathrm{TV}}(\mu,\nu). \tag{A6}$$

Combining this with Pinsker's inequality yields

$$W_1(\mu,\nu) \leq \mathrm{diam}(\mathcal{I})\sqrt{\frac{1}{2}D(\nu\|\mu)} = \sqrt{2B^2 D(\nu\|\mu)} \tag{A7}$$

and hence, μ is $T_1(B^2)$. Then, we can use the tensorization of the T_1 property [29, Proposition 3.4.4]

$$\mu_i = T_1(c) \text{ for all } i \Rightarrow \bigotimes_{i=1}^{n} \mu_i = T_1(nc) \tag{A8}$$

to conclude the stated Lemma. □

We now have all the tools at hand to present the proof:

Proof of Lemma 1. We start by defining a class of functions. Let $\sigma \in \{-1,1\}^T$ and define

$$F_\sigma(\hat{\boldsymbol{\xi}}) = \sum_{i=1}^{T} \sigma_i \frac{1}{N} \sum_{j=1}^{N} \hat{\xi}_{i,j}. \tag{A9}$$

We use this as a proxy for the ℓ_1-norm as for all realizations $\boldsymbol{\eta}$ there exists a choice σ_+ such that $\|\boldsymbol{\eta}\|_1 = F_{\sigma_+}(\boldsymbol{\xi})$. Also note that $\mathbb{E}[F_\sigma(\hat{\boldsymbol{\xi}})] = 0$ for all σ. The map F_σ is Lipschitz with respect to the ℓ_1-norm on $\mathcal{X}$ for all σ with Lipschitz constant

$$\|F_\sigma\|_{\text{Lip}} = \sup_{\boldsymbol{\xi}\neq\boldsymbol{\xi}'\in\mathcal{X}} \frac{|\sum_{i=1}^{T}\sigma_i \frac{1}{N}\sum_{j=1}^{N}\xi_{i,j} - \sum_{i=1}^{T}\sigma_i\frac{1}{N}\sum_{j=1}^{N}\xi'_{i,j}|}{\sum_{i=1}^{T}\sum_{j=1}^{T}|\xi_{i,j}-\xi'_{i,j}|} \tag{A10}$$

$$= \frac{1}{N}\sup_{\boldsymbol{\xi}\neq\boldsymbol{\xi}'\in\mathcal{X}} \frac{|\sum_{i=1}^{T}\sigma_i\sum_{j=1}^{N}(\xi_{i,j}-\xi'_{i,j})|}{\sum_{i=1}^{T}\sum_{j=1}^{T}|\xi_{i,j}-\xi'_{i,j}|} \tag{A11}$$

$$\leq \frac{1}{N}\sup_{\boldsymbol{\xi}\neq\boldsymbol{\xi}'\in\mathcal{X}} \frac{\sum_{i=1}^{T}\sum_{j=1}^{N}|\xi_{i,j}-\xi'_{i,j}|}{\sum_{i=1}^{T}\sum_{j=1}^{T}|\xi_{i,j}-\xi'_{i,j}|} \tag{A12}$$

$$= \frac{1}{N}, \tag{A13}$$

where we have applied the triangle inequality. Applying Theorem 1 to F_σ and the underlying distribution of $\hat{\boldsymbol{\xi}}$ yields

$$\mathbb{P}\left[F_\sigma(\hat{\boldsymbol{\xi}}) \geq \alpha\right] \leq \exp\left(-\frac{\alpha^2 N}{2TB^2}\right), \tag{A14}$$

irrespective of the particular choice of σ. We now split the parameter space $\boldsymbol{\xi} \in \mathcal{X}$ into parts $\mathcal{X}_\sigma$ such that

$$F_\sigma(\boldsymbol{\xi}) = \|\boldsymbol{\eta}\|_1 \text{ for all } \boldsymbol{\xi} \in \mathcal{X}_\sigma, \tag{A15}$$

and note that $\mathcal{X} = \bigcup_\sigma \mathcal{X}_\sigma$ and that $\mathcal{X}_\sigma \cap \mathcal{X}_{\sigma'}$ has measure zero for $\sigma \neq \sigma'$. With this, we can now conclude that

$$\mathbb{P}[\|\hat{\boldsymbol{\eta}}\|_1 \geq \alpha] = \mathbb{P}[\max_{\sigma'} F_{\sigma'}(\hat{\boldsymbol{\xi}})] \geq \alpha] \tag{A16}$$

$$= \sum_\sigma \mathbb{P}[\max_{\sigma'} F_{\sigma'}(\hat{\boldsymbol{\xi}})] \geq \alpha \cap \hat{\boldsymbol{\xi}} \in \mathcal{X}_\sigma] \tag{A17}$$

$$= \sum_\sigma \mathbb{P}[F_\sigma(\hat{\boldsymbol{\xi}}) \geq \alpha \cap \hat{\boldsymbol{\xi}} \in \mathcal{X}_\sigma] \tag{A18}$$

$$\leq \sum_\sigma \mathbb{P}[F_\sigma(\hat{\boldsymbol{\xi}}) \geq \alpha] \tag{A19}$$

$$\leq 2^T \exp\left(-\frac{\alpha^2 N}{2TB^2}\right). \tag{A20}$$

Bringing the prefactor into the exponent concludes the proof. □

To complete the guarantees for the reconstruction we need bounds on the largest and smallest non-zero eigenvalue of the matrix relevant for the reconstruction which is given by a Vandermonde matrix. We will make use of the fact that discrete Fourier transforms with equally spaced sampling points are ideally conditioned:

Lemma 3 (See Ref. [31])**.** *All singular values of the Vandermonde matrix $A \in \mathbb{C}^{T\times T}$ with entries $A_{j,k} = e^{-2\pi i \frac{j,k}{T}}$, $j,k \in \{0,1,\ldots,T-1\}$, are equal to $\sqrt{T}$.*

This Lemma underpins performance guarantees for a univariate discrete Fourier transform. If we have $x \in [0,2\pi)$ and $\omega_k \in \Omega = \{-\Omega_0,\ldots,\Omega_0\}$ we set $T = |\Omega| = 2\Omega_0 + 1$ and $x_j = 2\pi j/T$ so that

$$A'_{j,k} = e^{-ix_j\omega_k} = e^{-2\pi i\frac{j,k}{T}} e^{-2\pi\frac{j\Omega_0}{T}} \tag{A21}$$

reproduces the above formula up to a phase shift. This phase shift accounts for the difference of discrete Fourier transform and centered Fourier transform and does not alter the singular values as it can be seen as $A' = DA$ for a diagonal matrix with phases on the diagonal such that $D^\dagger D = \mathbb{I}$ and hence $A'^\dagger A' = A^\dagger D^\dagger D A = A^\dagger A$.

The multivariate extension of this is given by performing a DFT on a grid generated by the appropriate DFTs on each coordinate with the corresponding Vandermonde matrix being the tensor product of the individual Vandermonde matrices. This yields the following multivariate corollary of the above Lemma:

Corollary 1. *Consider a set of frequency vectors $\boldsymbol{\omega} \in \Omega$ such that the maximal frequency in every coordinate is $\omega_{\max}(i) = \max\{|\omega_i| \colon \boldsymbol{\omega} \in \Omega\}$. Then using a grid generated by choosing $T_i = 2\omega_{\max}(i) + 1$ equally spaced values in the interval $x_i \in [0, 2\pi)$ and performing the discrete Fourier transform on the product (over $T = \prod_i T_i$ points) yields a Vandermonde matrix A with all singular values equal to*

$$\sqrt{T} = \sqrt{\prod_{i=1}^{d} T_i} = \sqrt{\prod_{i=1}^{d} [2\omega_{\max}(i) + 1]}. \tag{A22}$$

With this we can now deliver the proof of Proposition 1 of the main text:

Proposition 2 (1)**.** *Let $\boldsymbol{c}_*$ be the vector of Fourier coefficients obtained by performing the protocol outlined in the main text with a number of samples per datapoint N. We can guarantee*

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x}) - g_{\boldsymbol{c}_*}(\boldsymbol{x})| \leq \epsilon\right] \geq 1 - \delta \tag{A23}$$

if

$$N = \frac{2\|M\|_\infty^2}{\epsilon^2}\left(\log\frac{1}{\delta} + T\log 2\right) \tag{A24}$$

and hence perform a total of

$$N_{\text{total}} = TN = \frac{2T\|M\|_\infty^2}{\epsilon^2}\left(\log\frac{1}{\delta} + T\log 2\right) \tag{A25}$$

invocations of the quantum learning model.

Proof. We first use the fact that

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x}) - g_{\boldsymbol{c}_*}(\boldsymbol{x})| = \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\sum_{\boldsymbol{\omega}\in\Omega} (c_{\boldsymbol{\omega}} - c_{*,\boldsymbol{\omega}}) e^{-i\boldsymbol{\omega}\boldsymbol{x}}\right| \tag{A26}$$

$$\leq \sum_{\boldsymbol{\omega}\in\Omega} \sup_{\boldsymbol{x}\in\mathcal{X}} |c_{\boldsymbol{\omega}} - c_{*,\boldsymbol{\omega}}||e^{-i\boldsymbol{\omega}\boldsymbol{x}}| \tag{A27}$$

$$= \|\boldsymbol{c} - \boldsymbol{c}_*\|_1. \tag{A28}$$

Note that this bound can not be improved without further assumptions on f and g.

We obtain $\boldsymbol{c}_*$ by applying a discrete Fourier transform to every data feature separately. Generically, we can write the relation between the outputs of a Fourier series $\boldsymbol{y} = \{f(\boldsymbol{x}_i)\}_{i=1}^T$ and the Fourier coefficients $\boldsymbol{c}$ as $\boldsymbol{y} = A\boldsymbol{c}$ where A is a Vandermonde-type matrix isomorphic to the tensor product of the local Vandermonde matrices associated to the discrete Fourier transforms on the different data features A_i. A particularly nice property of the discrete Fourier transform is that A is invertible in this case.

Note that experimentally, we use a sample mean estimate to approximate the output of the quantum model, hence we obtain an estimator $\hat{\boldsymbol{y}}$. Because of the linear structure, we can decompose our estimate into a perfect term and an error term

$$\hat{\boldsymbol{y}} = A\boldsymbol{c} + \hat{\boldsymbol{\eta}}. \tag{A29}$$

Our estimate for the underlying Fourier coefficients is then given by the least-squares estimator

$$\boldsymbol{c}_* = A^{-1}\hat{\boldsymbol{y}}, \tag{A30}$$

The ℓ_1-norm difference then becomes

$$\|\boldsymbol{c} - \boldsymbol{c}_*\|_1 = \|\boldsymbol{c} - A^{-1}(A\boldsymbol{c} + \hat{\boldsymbol{\eta}})\|_1 = \|A^{-1}\hat{\boldsymbol{\eta}}\|_1 = \frac{1}{\sqrt{T}}\|\hat{\boldsymbol{\eta}}\|_1, \tag{A31}$$

where we have exploited Corollary 1 and the fact that all singular values of A are equal to $\sqrt{T}$ which implies the last equality. To obtain a faithful estimate, we need to control the ℓ_1-norm of the estimation error. To do so, we note that

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{X}}|f(\boldsymbol{x}) - g_{\boldsymbol{c}_*}(\boldsymbol{x})| \geq \epsilon\right] \leq \mathbb{P}\left[\|A^{-1}\boldsymbol{\eta}\|_1 \geq \epsilon\right]. \tag{A32}$$

as the latter event implies the other. As

$$\mathbb{P}\left[\|A^{-1}\boldsymbol{\eta}\|_1 \geq \epsilon\right] = \mathbb{P}\left[\|\boldsymbol{\eta}\|_1 \geq \sqrt{T}\epsilon\right], \tag{A33}$$

we can now apply Lemma 1 to the random variable $\hat{\boldsymbol{\eta}}$ with $\alpha = \sqrt{T}\epsilon$, $B = \|M\|_\infty$ to obtain

$$\mathbb{P}\left[\sup_{\boldsymbol{x}\in\mathcal{X}}|f(\boldsymbol{x}) - g_{\boldsymbol{c}_*}(\boldsymbol{x})| \geq \epsilon\right] \leq \exp\left(\log(2)T - \frac{\epsilon^2 N}{2\|M\|_\infty^2}\right). \tag{A34}$$

Setting the right hand side equal to δ and solving for N yields the statement of the Proposition. □

Appendix B: Sample complexity of estimating multiple observables

Obtaining the outputs of a quantum learning model defined via expectation values also comes with a overhead. For the sake of completeness, we give a proof of the well-known sample complexity, compare e.g. Ref. [32].

Lemma 4. *Let $\hat{\boldsymbol{\xi}} = (\hat{\xi}_{i,j})_{i=1\,j=1}^{T\quad N} \in \mathcal{X}$ be a collection of i.i.d. zero-centered random variables such that $|\hat{\xi}_{i,j}| \leq B$ for all i and j. Let $\hat{\boldsymbol{\eta}}$ denote the vector of the T sample means $\hat{\eta}_i = \frac{1}{N}\sum_{j=1}^N \hat{\xi}_{i,j}$. We can guarantee*

$$\mathbb{P}\left[\|\hat{\boldsymbol{\eta}}\|_\infty \geq \epsilon\right] \leq \delta \tag{B1}$$

for a total number of i.i.d. copies

$$N_{\text{total}} = NT \geq \frac{2B^2}{\epsilon^2} T \log\frac{2T}{\delta}. \tag{B2}$$

Proof. For every entry of $\hat{\boldsymbol{\eta}}$, we have by Hoeffding's inequality that

$$\mathbb{P}\left[|\hat{\eta}_i| \geq \epsilon\right] \leq 2\exp\left(-\frac{N\epsilon^2}{2B^2}\right). \tag{B3}$$

We can then use the union bound to conclude

$$\begin{aligned}
\mathbb{P}\left[\|\hat{\boldsymbol{\eta}}\|_\infty \geq \epsilon\right] &= \mathbb{P}\left[\bigcup_{i=1}^T \{|\hat{\eta}_i| \geq \epsilon\}\right] && \text{(B4)}\\
&\leq \sum_{i=1}^T \mathbb{P}\left[|\hat{\eta}_i| \geq \epsilon\right] && \text{(B5)}\\
&= \sum_{i=1}^T 2\exp\left(-\frac{N\epsilon^2}{2B^2}\right) && \text{(B6)}\\
&= 2T\exp\left(-\frac{N\epsilon^2}{2B^2}\right). && \text{(B7)}
\end{aligned}$$

Equating the right hand side to δ and solving for N yields the claim of the Lemma. □

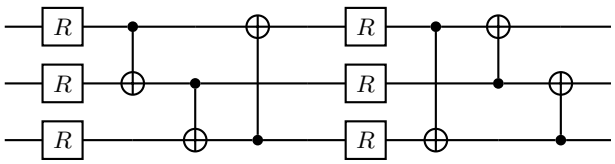

Figure 3. Example implementation of a trainable circuit block $W^{(l)}$ for three qubits with two block layers. The R-gates are arbitrary rotation gates of the form $R(\alpha,\beta,\gamma) = R_X(\alpha)R_Z(\beta)R_X(\gamma)$. In the b-th block layers, the CNOT gates connect i and $(i+b) \mod n$, where n is the number of qubits.

Appendix C: Discussion of ill-conditioning

When working with datasets, one does not have control over the datapoints for which we have corresponding labels, leading to a potentially ill-conditioned situation. This may arise when data points are too close to each other. To illustrate this, consider the case where the Fourier matrix A is square. The condition number of A, $\kappa(A)$, is given as the ratio of largest and smallest singular value. Since $\kappa(A) = \kappa(A^T)$, we can consider A^T instead of A, such that the i'th column of A^T, $\boldsymbol{a}_i$, is associated with data point $\boldsymbol{x}_i$. The variational formulation of singular values gives for the condition number

$$\kappa(A^T) = \frac{\max_{\|\boldsymbol{z}_1\|_2=1} \left\|A^T\boldsymbol{z}_1\right\|_2}{\min_{\|\boldsymbol{z}_2\|_2=1} \left\|A^T\boldsymbol{z}_2\right\|_2}. \tag{C1}$$

Choosing $\boldsymbol{z}_1 = (\boldsymbol{e}_i + \boldsymbol{e}_j)/\sqrt{2}$ and $\boldsymbol{z}_2 = (\boldsymbol{e}_i - \boldsymbol{e}_j)/\sqrt{2}$, where $\boldsymbol{e}_i$ is the i-th standard basis vector, yields the lower bound

$$\kappa(A^T) \geq \frac{\left\|\boldsymbol{a}_i + \boldsymbol{a}_j\right\|_2}{\left\|\boldsymbol{a}_i - \boldsymbol{a}_j\right\|_2}, \tag{C2}$$

which clearly blows up when two column vectors approach each other, which happens exactly when two datapoints are too close.

Appendix D: Additional information on numerical experiments

In this section, we collect further information and interpretation for the numerical experiments.

1. Example of a trainable block

An example of a trainable block $W^{(l)}$ of the *Strongly Entangling Layer* type discussed in the main text for three qubits is shown in Fig. 3.

2. Loss curves for numerical examples

Fig. 4 shows the loss curves for the three examples discussed in the main text. For completeness, we show the results from Fig. 2 in the first row again.

3. Example of random parametrized quantum circuits

As a further numerical example, we use a synthetic dataset which is "natural" for the quantum models we consider. The learning problem is to predict the output of a randomly initialized re-uploading model of the structure introduced at the beginning of this section with input dimension $\boldsymbol{x}_j \in \mathbb{R}^4$, $L = 2$ and $B = 2$. We compute $N = 3500$ random samples.

In the second row of Fig. 4, training trajectories are shown for two quantum learning models with $L = 3$ and $B \in \{1, 3\}$ and the classical surrogate for this problem. Here, we observe that even the smallest quantum learning model we analyze which has only one block layer has very smooth training curves. We observe that adding more block layers improves the performance of the quantum learning model, but it is always significantly less accurate than the classical surrogate. The good performance of the classical surrogate is completely expected as the frequency structure of the generating process is contained in the frequency

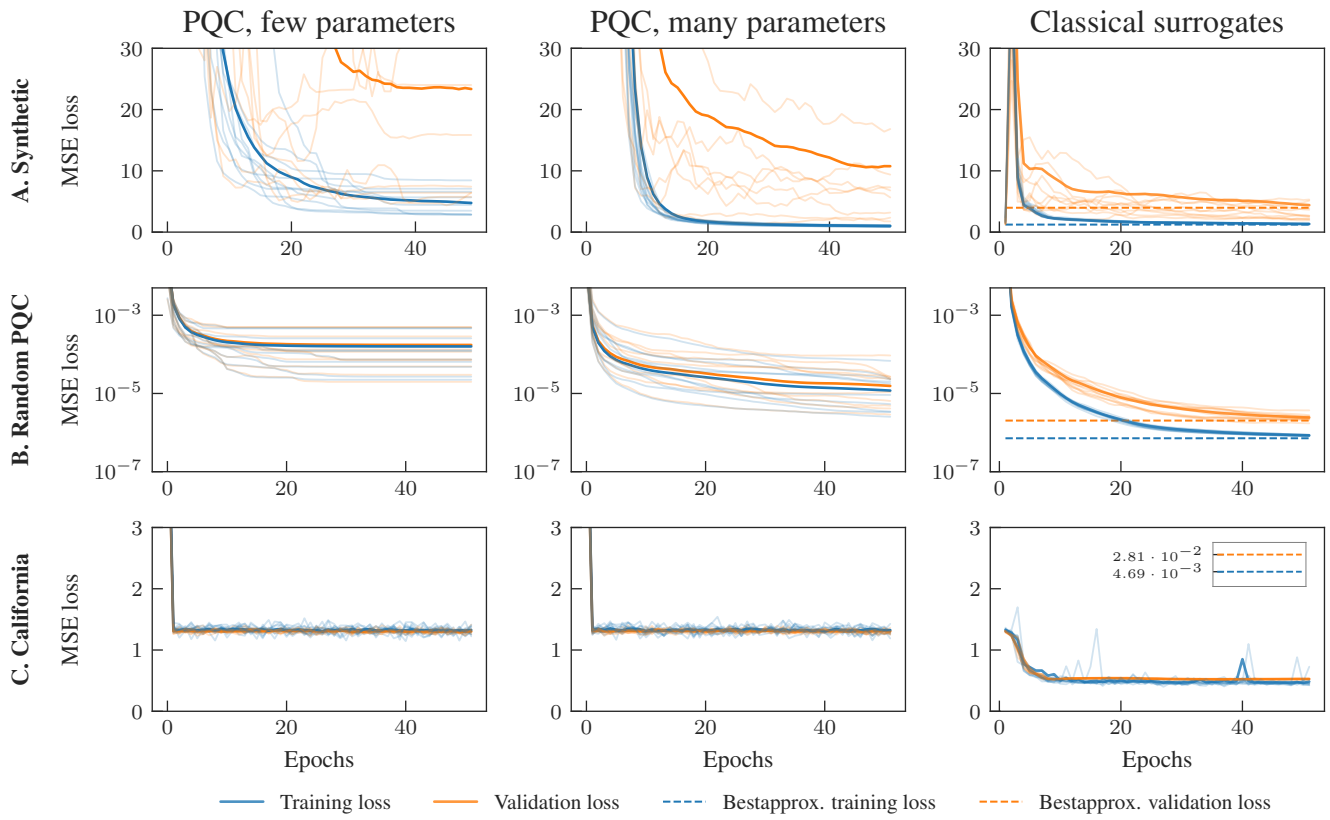

Figure 4. (**Row-wise**) In the top row the synthetic dataset (models have $L = 2$), in the middle row the dataset sampled from a randomly initialized PQC (with $L = 2$, $B = 2$, the models have $L = 3$) and in the bottom row the California housing dataset (models have $L = 3$). (**Column-wise**) In the left column are quantum models where the number of block layers is $B = 1$, *i.e.* the number of trainable parameters $\boldsymbol{\theta}$ is low. Depicted in the middle row are quantum models with $B = 3$, resulting in a higher number of trainable parameters $\boldsymbol{\theta}$. In the right column are the corresponding classical models. For the two smaller datasets the best approximation to the training data was computed by directly solving the linear least square problem. Note that the "best approximation validation loss" line only gives the validation loss corresponding to the lowest possible training loss, not the lowest possible validation loss. (**General**) All loss curves of the quantum models are the averages $N = 50$ training runs, each with randomly initialized weights (darker color). For each model, ten individual runs were plotted for illustration.

structure of the surrogate. Consequently, the optimization of the classical surrogate in this instance is more or less equal to the surrogation process introduced above, except that we did not have control over the data-points in this case which is alleviated by the fact that we have sufficiently many of them. Still, as this also holds true for the frequency structure of the quantum models, this results underlines how the convex loss landscape of the classical surrogate can lead to favorable trainability properties compared to the much more complex loss landscape of the quantum models.

All in all, we see that the observations made for the first dataset also hold true in this setting. As we increase the number of parameters for the quantum models, we see a transition from a rugged loss landscape (high standard deviation in the loss functions) towards increasingly behaving like their classical surrogates (again accompanied by a drop in standard deviation of the loss functions). Despite the fact that we chose a very natural problem for the quantum models in question, they are consistently outperformed by their classical surrogates.

4. Example of the California Housing Dataset

The California housing dataset is a canonical small benchmark regression problem from classical machine learning. The dataset consists of $N = 20640$ samples with input dimension $\boldsymbol{x}_j \in \mathbb{R}^8$. The task is to predict the value of houses in the price range of $\$15000 - \$500000$ which we map to the interval $0.15 - 5$.

Looking at the third row of Fig. 4, we observe that even for models with very few parameters, we are immediately drawn

into a local minimum for both parameter counts of the quantum learning model. We assume this is due to unfavorable loss landscapes for the quantum learning models for this particular dataset combined with the limited expressivity of the models in terms of how many Fourier coefficients were available. The classical surrogate, on the contrary, manages to reach an improved validation and training loss. It is curious that the training loss of the classical surrogate does not approach the value resulting from direct inversion of the problem which is indicated by dashed lines in the inset. This is likely due to numerical reasons resulting from the LBFGS solver we employ. Altogether, while we are not able to observe a transition from high variance and rugged loss landscape to lower variance regimes as with the other models, the classical surrogate still reaches better results than the corresponding quantum models.

Appendix E: Author Contributions

F.J.S. conceived and conducted the numerical experiments. J.J.M. envisioned the theoretical part of this work with support from F.J.S. J.E. supported research and development. All authors contributed to the writing of the manuscript.

Appendix F: Data Availability

Code for implementations and data of the numerical experiments conducted in this work will be made available upon reasonable request.