Opportunities and limitations of explaining quantum machine learning

Elies Gil-Fuster,[1,2] Jonas R. Naujoks,[2] Wojciech Samek,[2,3,4] and Jens Eisert[1,2,5]

[1]*Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, 14195 Berlin, Germany*
[2]*Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany*
[3]*Department of Electrical Engineering and Computer Science, Technische Universität Berlin, 14109 Berlin, Germany*
[4]*BIFOLD – Berlin Institute for the Foundations of Learning and Data, 14109 Berlin, Germany*
[5]*Helmholtz-Zentrum Berlin für Materialien und Energie, 14109 Berlin, Germany*

Introduction. State-of-the-art Machine Learning (ML) models have shown great results across a wide range of applications [1, 2]. From image recognition [3] to natural language processing [4], though, deep Neural Networks (NNs) display one negative property: their inner working is difficult for humans to interpret [5]. Questions like *what is the role of each parameter in the network?*, or *how relevant is each input dimension for the produced output?* occupy the field of eXplainable Artificial Intelligence (XAI) [6]. In the advent of Quantum Machine Learning (QML) [7–9], where access to quantum computers promises to improve the performance of large-scale learning tasks, the story is no different when it comes to explainability. In this work, we offer a preview into the upcoming field of eXplainability in Quantum Machine Learning (XQML) [10–12], comprising techniques to explain the behavior of QML models.

Results. Our first contribution is a didactic presentation of explainability methodologies in ML, discussing their key components and their direct transferrability to QML models based on Parametrized Quantum Circuits (PQCs). We discuss how the fundamental differences between NNs and PQCs lead to incompatibilities between some classical explainability techniques and QML models, like the need to store intermediate states or clone information. We identify the desired features that upcoming XQML algorithms should fulfill, and so characterize the shortcomings of just applying classical methods off-the-shelf. As most classical XAI algorithms to date are specific to NNs, we are a priori left with only black-box or model-agnostic techniques for QML. Our second contribution resolves this problem, in that we propose two model-specific XQML techniques specifically for PQC-based QML models. We analyze the model-specific techniques we introduced by performing numerical experiments on a synthetic learning task. We observe that the XQML algorithms we propose work as desired, and in doing so we focus on the important aspects of benchmarking XQML techniques for future works.

Main ideas. The umbrella term XAI addresses different aspects of *understanding what a learning model does*. For example, we might want to understand the behavior of a model over an entire data domain, or we might be interested in what caused a particular pattern to be assigned a particular class. In this work, we focus on the latter: we propose algorithms which take as input a trained QML model and a single pattern, and return a relevance score for each of the pattern's input dimensions. Building on expressivity results for PQCs [13], we combine the Taylor and Fourier representation of output functions to assign relevance scores to each input component based on all their single-component derivatives [14]. We also exploit the kernel picture for quantum models to propose a layer-wise propagation method inspired on those introduced for NNs [15].

Potential impact. This submission ought to be seen as a guide for what is to come in the field of XQML [10–12]. We combine a detailed presentation of the relevant concepts, a proposal for PQC-specific XQML techniques, as well as illustrative numerical experiments, in a way that we hope will serve as guide for future progress in this field. The potential impact of this work would be therefore to positively affect the rate of progress in explainable QML in the near future and beyond. Explainability in NNs started to be studied mostly after the surge in performance. In QML we have the opportunity to let considerations of explainability guide our design of PQCs to ensure that they remain explainable throughout and so dodge difficult dilemmas of performance versus fairness in tasks with wider societal impact [16].

In a broader context. We frame this submission within the quest for *good* design principles for PQCs for QML. Next to the usually considered properties of expressivity, generalization, trainability (and non-dequantization), we believe explainability should also be taken into account. This way, developing XQML techniques contributes to identifying how PQCs behave in QML tasks, and can guide us toward PQC Ansätze that can achieve quantum advantage. Also, XQML techniques can be extended beyond learning tasks to understand other aspects of the inner working of PQCs. For example, Ref. [11] already pointed at an application of XAI to characterize the importance of each gate in a circuit towards the output state. Similar ideas can be employed with the aim of *prunning* a PQC, resulting in smaller, shallower circuits that can be implemented more reliably on current hardware.

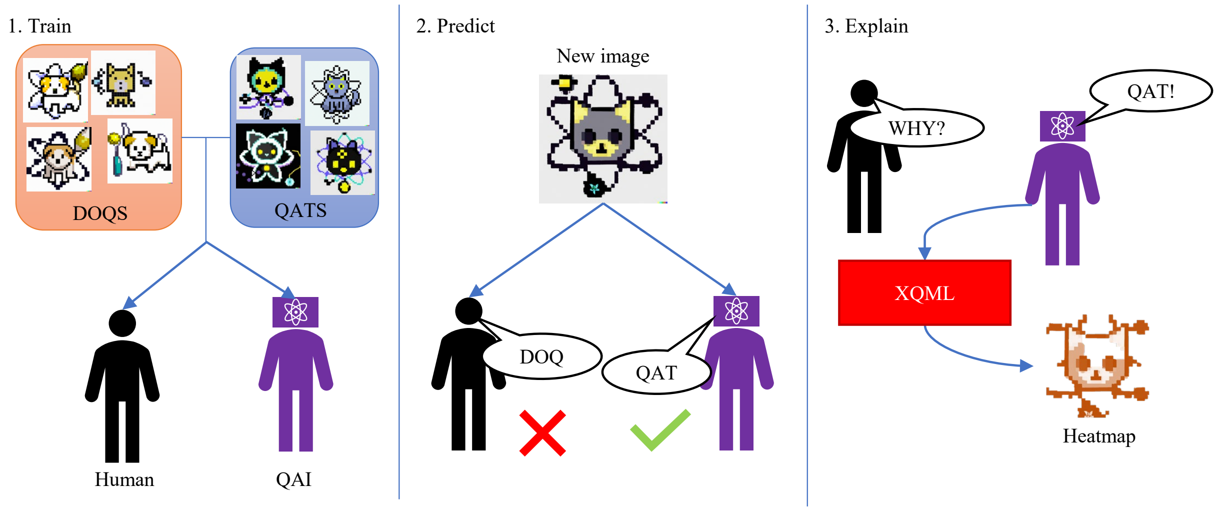

Figure 1. Sketch depicting the pipeline of post-hoc eXplainability in Quantum Machine Learning (XQML). After the usual training and predicting phase, there is a third step: explanation. In XQML we develop algorithms which, given a trained model and an input, return a relevance heatmap for the predicted class for the given input pattern.

From here on, we shortly present the necessary tools to discuss explainability in variational QML.

I. Flavors of explainability. The study of explainability can be divided across different axes [6]. For example, ante-hoc explainability deals with the study of learning models that are designed to be intrinsically interpretable or explainable; whereas post-hoc explainability considers the model resulting after the training step and aims to recover its potentially obscure internal reasoning in a humanly-understandable way. Another key distinction is to be made between *global* and *local* explanations: while global explanations characterize the role of the building blocks of the model across the entire data domain; local explanations in turn ask which parts of a single input pattern were more relevant for the model to produce its output. Finally, *model-agnostic* algorithms differentiate themselves from *model-specific* ones in that they do not require access to the inner workings of the functions, but rather they need only evaluate the functions themselves.

An illustrative example are most physics theories, which lead to ante-hoc global explanations, since the parameters of a physical theory usually correspond to human-interpretable quantities, like friction coefficients, coupling strengths, etc. Conversely, in this work we focus on post-hoc local explanations, as depicted in Fig. 1. In post-hoc local explainability one might consider a neural network trained to recognize images of cats and dogs, and a single picture of a cat. The goal of the explanation algorithm is then to highlight which specific pixels in the image are responsible for the model's prediction that the image is indeed a cat. This flavor of explainability can help for instance in detecting the "clever Hans" effect in pattern recognition [17].

II. Relevance score as a heatmap. We can think of a trained ML model as a function $f(x)$ that quantifies a certain property (like the presence of an object in an image). Our goal is to associate a relevance score $R_i(x)$ to each data component x_i. The collection of relevance values for all pixels in the image is called a relevance *heatmap* $R(x)$, as it has the same dimensions as x. A heatmap $R(x)$ is *conservative* if its sum equals the original function $f(x) = \sum_i R_i(x)$, for all x. Conservativity is a favorable property as it results in a form of normalization that makes heatmaps more human interpretable [15].

III. Taylor decomposition relevance score. A common starting point in XAI is to truncate the Taylor expansion of the model's function to the first order: $f(x) = f(\tilde{x}) + (\nabla_x f(\tilde{x}))^\intercal (x - \tilde{x}) + \varepsilon$ [14]. By choosing the base point of the Taylor expansion to be such that $f(\tilde{x}) \approx 0$ (which we call a *root point*), and assuming that the contribution of the higher-order terms is negligible, we reach $f(x) \approx \sum_i \partial_i f(\tilde{x})(x_i - \tilde{x}_i)$. From here, we can define a heatmap based on this Taylor expansion $R_i(x) := \partial_i f(\tilde{x})(x_i - \tilde{x}_i)$. The assumption that only the first-order terms dominate the sum consequently results in a conservative and positive heatmap. Note that this relevance score is black box model-agnostic: provided the root point $\tilde{x}$, we need only evaluate the derivatives of the function. Qualitatively, a good root point should be similar to the actual data point x to sufficiently contextualize the explanation, which means it should minimally deviate from it [15]. In general, though, and especially for non-linear functions, it might be difficult to find a root point $\tilde{x}$ for which the higher-order terms are sufficiently small'. So while this Taylor decomposition offers an illustrative starting point, we must consider improved techniques.

IV. Taylor-Fourier relevance score for PQCs We propose a specialization of this relevance score to PQCs, by using the well-known Fourier picture [13] of PQCs. We recognize that, for PQCs where input data is encoded only once, the approximation quality can easily be improved by taking all-order derivatives with respect to every single component, instead of only the first derivatives. To this end, we consider a different grouping of the terms in the Taylor series $f(x) = f(\tilde{x}) + \sum_i \sum_{k=1}^{\infty} \partial_i^k f(\tilde{x})(x_i - \tilde{x}_i)^k/k! + \varepsilon$, where now ε contains only the cross-derivative terms. Given a root point $\tilde{x}$ fulfilling $f(\tilde{x}) \approx 0$ and the cross terms ε being small, we have $f(x) \approx \sum_i \sum_{k=1}^{\infty} \partial_i^k f(\tilde{x})(x_i - \tilde{x}_i)^k/(k!)$. In particular, for f, a degree-1 trigonometric polynomial in several variables, we show that $\sum_{k=1}^{\infty} \partial_i^k f(\tilde{x})(x_i - \tilde{x}_i)^k/(k!) = \sin(x_i - \tilde{x}_i)\partial_i f(\tilde{x}) + (1 - \cos(x_i - \tilde{x}_i))\partial_i^2 f(\tilde{x})$. With this, we reach a relevance score specific to trigonometric polynomials: $R_i(x) = \sin(x_i - \tilde{x}_i)\partial_i f(\tilde{x}) + (1 - \cos(x_i - \tilde{x}_i))\partial_i^2 f(\tilde{x})$. This relevance score represents the first model-aware explanation technique for QML. On the one hand, we only need black-box access to the PQC function $f(x)$ in order to compute the relevance scores. On the other hand, we do not run into exponential memory requirements from needing to store intermediate states. This idea can be generalized to PQCs with data re-uploading, which we leave for future work.

V. A model-specific relevance score for PQCs. We would like a PQC-specific explanation technique that exploits the inner structure of the PQC. Our second novel technique is called *Quantum Layerwise Relevance Propagation* (QLRP) as an analogous to the well-established technique of *layerwise relevance propagation* (LRP) for NNs [15]. Our technique treats an encoding-first PQC as a two-step process: first, prepare a data-dependent quantum state $\rho(x)$, and second, measure a task-dependent observable $\mathcal{M}$. The algorithm then propagates the relevance in a layer-wise way in this two step process in reverse: (1) from the output function $f(x) = \mathrm{tr}(\rho(x)\mathcal{M})$, we first assign a score function to each entry $\rho_{jk}(x)$: $f(x) \mapsto R_{jk}(\rho(x))$; (2) We distribute the relevance of each entry $\rho_{jk}(x)$ onto each input dimension x_i: $R_{jk}(x) \mapsto R_i^{jk}(x)$. In order to obtain the total relevance of each entry, we simply need to sum over all the entries of the quantum state: $R_i(x) = \sum_{j,k} R_i^{jk}(x)$.

Our implementation of QLRP employs a full simulation of the quantum circuit as a digital twin neural network (twiNN) to exploit the information present in the intermediate data-dependent state. In this case, it suffices for us to consider a twiNN with a single hidden layer. In the first layer, the twiNN computes a real-valued data-dependent matrix that represents the density matrix of the data-dependent state. In the second and final layer, the twiNN computes the expectation value of the task-dependent observable on the data-dependent state as an inner product of matrices –that is, as a linear output layer. Albeit having some practical limitations, this explanation technique represents the first model-specific application for quantum learning models, and the first quantum version of an LRP technique.

VI. Numerical experiment. We apply our newly proposed XQML techniques in a simple learning task, and confirm that QLRP performs better than the black box Taylor-Fourier method, and similarly well to other standard XAI techniques. Our experiments are not geared toward showing an advantage, but rather to display the different properties of different explanation techniques. At the time of submission we have not finished running experiments, so we can only show partial results in Fig. 2.

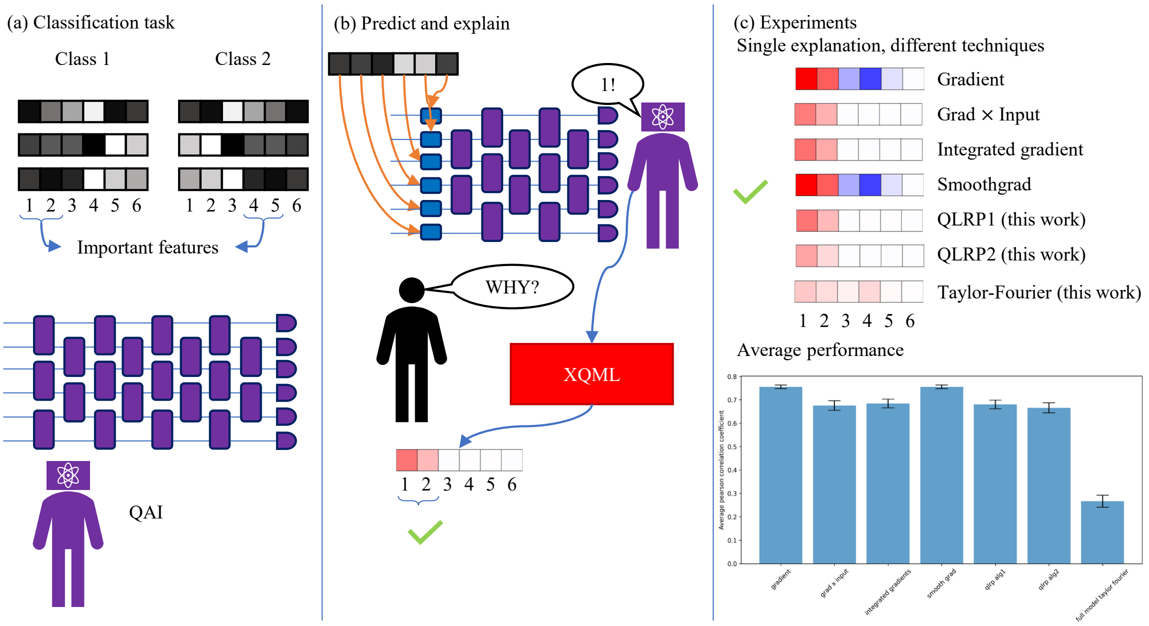

Figure 2. Sketch of the numerical experiments. After training a PQC, we generate explanations using several XQML techniques, both known ones from the classical XAI literature and novel ones. We show how the different techniques produce slightly different relevance heatmaps.

[1] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, in *Neural networks: Tricks of the trade* (Springer, Berlin, 2012), pp. 9–48.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT Press, Cambridge (MA), 2016).

[3] K. He, X. Zhang, S. Ren, and J. Sun, Proc. IEEE Conf. Comp. Vision and Pattern Rec. (CVPR) pp. 770–778 (2016).

[4] L. Deng, G. Hinton, and B. Kingsbury, IEEE Int. Conf. Ac., Speech and Signal Proc. (ICASSP) pp. 8599–8603 (2013).

[5] S. Bach, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, *Analyzing classifiers: Fisher vectors and deep neural networks* (2015), 1512.00172.

[6] W. Samek, T. Wiegand, and K.-R. Müller (2017), arXiv:1708.08296.

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

[8] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019).

[9] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New J. Phys. **18**, 023023 (2016).

[10] P. Steinmüller, T. Schulz, F. Graf, and D. Herr, *explainable ai for quantum machine learning* (2022), 2211.01441.

[11] R. Heese, T. Gerlach, S. Mücke, S. Müller, M. Jakobs, and N. Piatkowski, *Explaining quantum circuits with shapley values: Towards explainable quantum machine learning* (2023), 2301.09138.

[12] L. Pira and C. Ferrie, *On the interpretability of quantum neural networks* (2024), 2308.11098.

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

[14] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, Pattern Recognition **65**, 211 (2017), ISSN 0031-3203.

[15] G. Montavon, W. Samek, and K.-R. Müller, Digital Signal Processing **73**, 1 (2018), ISSN 1051-2004.

[16] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (2015), pp. 1721–1730.

[17] J. Kauffmann, L. Ruff, G. Montavon, and K.-R. Müller, *The clever hans effect in anomaly detection* (2020), 2006.10609.