Noisy quantum kernel machines

Valentin Heyraud◉, Zejian Li◉, Zakari Denis, Alexandre Le Boité◉, and Cristiano Ciuti◉

Université Paris Cité, CNRS, Matériaux et Phénomènes Quantiques, F-75013 Paris, France

In the noisy intermediate-scale quantum era, an important goal is the conception of implementable algorithms that exploit the rich dynamics of quantum systems and the high dimensionality of the underlying Hilbert spaces to perform tasks while prescinding from noise-proof physical systems. An emerging class of quantum learning machines is that based on the paradigm of quantum kernels. Here, we study how dissipation and decoherence affect their performance. We address this issue by investigating the expressivity and the generalization capacity of these models within the framework of kernel theory. We introduce and study the effective kernel rank, a figure of merit that quantifies the number of independent features a noisy quantum kernel is able to extract from input data. Moreover, we derive an upper bound on the generalization error of the model that involves the average purity of the encoded states. Thereby we show that decoherence and dissipation can be seen as an implicit regularization for quantum kernel machines. As an illustrative example, we report exact finite-size simulations of machines based on chains of driven-dissipative quantum spins to perform a classification task, where the input data are encoded into the driving fields and the quantum physical system is fixed. We determine how the performance of noisy kernel machines scales with the number of nodes (chain sites) as a function of decoherence and examine the effect of imperfect measurements.

I. INTRODUCTION

In recent years, machine learning has blossomed in a wide variety of fields and delivered a large number of applications driven by the achievements of the ever-developing field of artificial neural networks, particularly those presenting deep architectures [1,2]. Neural networks build predictions upon processing the input data of interest through a series of parametrized nonlinear transformations, the (typically numerous) parameters of which are determined by training. This optimization procedure most often consists in minimizing a task-dependent loss function that quantifies the error of the parametrized model over a training dataset. This is most often implemented via software executed on standard computers. As a result, the growing demand for computational resources and energy for training such deep architectures on ever-increasing amounts of data makes its long-term sustainability uncertain [3]. In this context, devolving computationally demanding tasks to machine-learning devices with suitable physical systems acting as hardware is emerging as a relevant alternative. However, while the neural-network sequential architecture is well suited for software implementations on standard computers, the great number of parameters to be tuned during training remains in practice an obstacle to physical implementations. A simpler alternative approach is provided by the category of "shallow models," such as reservoir computing [4] or extreme learning machines [5], which have led to physical proposals [6,7] and experimental realizations [8,9]. In such machines, the input data are encoded in the dynamics of a physical system and the associated predictions are obtained by considering a linear combination of measured observables, weighted by a set of trainable parameters to be optimized by training. Importantly, this is done while keeping the parameters of the physical system fixed, hence requiring hardly any degree of control over the system. Kernel machines, the trial functions of which can be represented in terms of positive semidefinite and symmetric kernel functions [10], belong to this category. More generally, kernel theory has proved to be a very useful tool to understand a wide range of machine-learning algorithms. Recently, a close connection between kernel machines and deep neural networks in the infinite width limit has been established [11], further extending the relevance of these methods. In parallel to the advent of quantum information, the last decade has also witnessed a growing interest in the emerging field of quantum machine learning [12,13], a research domain that explores the potential advantages of quantum systems for machine-learning applications. Due to the success of deep neural-network algorithms, a large amount of work in this field has been devoted to finding quantum analogs to neural-network models [14], and more generally to finding brain-inspired algorithms to be implemented on quantum devices [15]. Parametrized quantum circuits used as trainable *Ansätze* [16] appeared as natural candidates for such a generalization. These models, often called quantum neural networks [17], are among the most studied quantum machine-learning models and significant progress has been achieved in the comprehension of their properties. Analogously to classical systems, quantum "shallow" machines have also been put forward, such as those based on quantum reservoir computing, extreme learning, and quantum kernels [18–29], where the physical system (the network) is fixed and the optimization concerns only a linear map acting on measured outcomes.

Most often, quantum machine-learning investigations have been focusing on isolated quantum systems with unitary dynamics. At present, however, we are in the so-called noisy intermediate-scale quantum era [30]: most quantum devices

within practical reach are subject to a significant degree of dissipation and/or decoherence. An important problem is therefore to understand the impact of realistic noise on quantum machine-learning settings. The literature on the subject is yet in its infancy. For time-dependent tasks, one study on quantum reservoir computing suggested that dissipation increases the processing capacity and the nonlinearity of the embedding, at the price of a reduced memory capacity of the system [31]. An advantageous scaling of the performance of a quantum reservoir-computing scheme, as compared to its classical counterpart, was recently reported [32]. However, a systematic study of the dissipation and decoherence on quantum machine-learning models is missing. In particular, to the best of our knowledge, no investigation has explored its role on the important class of quantum kernel machines. In this paper, we investigate the use of open quantum systems as noisy quantum kernel machines. Within the formalism of kernel theory, we show how the expressive power and generalization capacity of the corresponding nonlinear feature maps are controlled by both the dissipation and decoherence affecting the system as well as the level of experimental uncertainty on the physical measurements. We introduce and study the effective kernel rank to quantify the effective number of independent features a noisy quantum kernel is able to extract from the input data. Moreover, we derive an upper bound on the generalization error of the model that involves the average purity of the encoded states. As an illustrative example, we simulate noisy quantum kernel machines implemented via driven-dissipative chains of spins. We provide a comprehensive study of the performance of noisy quantum kernel machines, showing how they scale with the number of network nodes (chain sites) and the degree of dissipation and decoherence. The paper is organized as follows. In Sec. II, we describe the general scheme for encoding the input data into the quantum system dynamics and decoding the output through measurements. In Sec. III, we analyze the noisy quantum kernel machine within the kernel-theory framework. In particular we study the link between the kernel spectrum and important properties of machine-learning models, such as the expressive power and the generalization capacity. We introduce and study the effective kernel rank. Within a statistical-learning approach, we provide an upper bound on the generalization error for noisy quantum kernels. In Sec. IV, we describe a class of noisy quantum kernel machines based on driven-dissipative chains of spins. We report a comprehensive study of the dependence of the performance metrics on the system size and noise for this class of models in Sec. V. Finally, conclusions and perspectives are drawn in Sec. VI. The most technical details are reported in Appendices A–C.

II. GENERAL SCHEME

The objective of supervised learning is to approximate a causal relation between elements $\boldsymbol{x}$ of an input set $\mathcal{X}$ and some target quantities $y \in \mathcal{Y}$, based upon a set of known training examples $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \,|\, i = 1, \ldots, N_{\text{train}}\}$. The input features are considered as independent realizations of a random variable following a probability distribution $p(x)$ on $\mathcal{X}$. In the following, we denote $\mathbb{E}_p[f(\boldsymbol{x})]$ the expectation value of a quantity $f(\boldsymbol{x})$ over the distribution p [33]. We also define the

corresponding centered quantity as

$$\delta f(\boldsymbol{x}) = f(\boldsymbol{x}) - \mathbb{E}_p[f]. \tag{1}$$

Upon assuming the inputs and target quantities are related according to an unknown ground-truth function $y_i = y(\boldsymbol{x}_i)$, we aim to approximate it using a trial function f parametrized by $\boldsymbol{w}$, to be optimized using the training set $\mathcal{S}$. The specific form of f depends on the considered model architecture. In this paper, we describe noisy quantum kernel machines exploiting the dynamics of open quantum systems to generate such a trial function. This scheme is summarized pictorially in Fig. 1.

A. Encoding on the quantum system

Let us consider a system initially prepared in a state $\hat{\rho}_0$. For each element of the input space, represented by a vector $\boldsymbol{x} \in \mathcal{X}$, a procedure can be defined to encode it into the nonunitary dynamics of a generic open quantum system. As will be shown in Sec. IV, this can be achieved, for instance, by encoding the input vector in a proper modulation of the driving fields acting on the system. We consider the dynamics of the open quantum system to be described by a Lindblad master equation [34] of the form

$$\frac{\partial \hat{\rho}}{\partial t} = -\frac{i}{\hbar}[\hat{H}, \hat{\rho}] + \sum_{j=1}^{N} \gamma_j \mathcal{D}(\hat{A}^j)[\hat{\rho}], \tag{2}$$

where γ_j is the relaxation rate at site i and the dissipator $\mathcal{D}(\hat{A})$ denotes the superoperator:

$$\mathcal{D}(\hat{A})[\hat{\rho}] = \hat{A}^\dagger \hat{\rho} \hat{A} - \tfrac{1}{2}\{\hat{A}^\dagger \hat{A}, \hat{\rho}\}. \tag{3}$$

Note that the Lindblad operator $\hat{A}^j$ depends on the considered system-bath interaction. The master equation describes the evolution from an initial density matrix into a final density matrix:

$$\hat{\rho}(\boldsymbol{x}, t) = \mathcal{M}(\boldsymbol{x}, t)[\hat{\rho}_0], \tag{4}$$

where the completely positive trace-preserving map $\mathcal{M}(\boldsymbol{x}, t)$ is the propagator of the Lindblad master equation capturing the nonunitary evolution of $\hat{\rho}_0$. It depends on $\boldsymbol{x}$ via the encoding procedure: If the input is encoded in driving fields, as we will consider later, the Hamiltonian, and consequently the density matrix at any time, bears a dependence on the input. In principle, one could also encode the input into a modulation of the loss rates, although we will not treat this case here. In what follows, when considering a fixed final time t_f for the time evolution, we denote $\mathcal{M}(\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}, t_f)$ and $\hat{\rho}(\boldsymbol{x}) = \hat{\rho}(\boldsymbol{x}, t_f)$ to simplify the notation.

B. Decoding through measurements

At time t_f, after the encoding procedure, we extract the processed information by performing a set of measurements of the system. Given the density matrix $\hat{\rho}(\boldsymbol{x})$ and a set of system observables $\mathcal{O} = \{\hat{O}_j \,|\, j = 1, \ldots, P\}$, information about the response of the open quantum system to the input $\boldsymbol{x}$ is contained in the following vector:

$$\boldsymbol{\phi}(\boldsymbol{x}) \equiv (1, \langle \hat{O}_1 \rangle_{\boldsymbol{x}}, \ldots, \langle \hat{O}_P \rangle_{\boldsymbol{x}})^T, \tag{5}$$
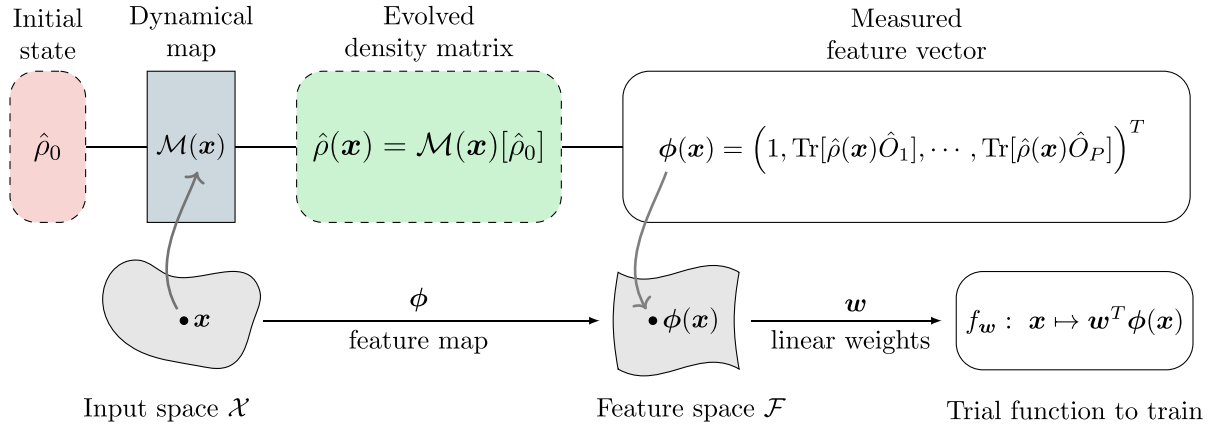
FIG. 1. Scheme of a noisy quantum kernel machine. An element $\boldsymbol{x}$ of the input space $\mathcal{X}$ is encoded into a density matrix $\hat{\rho}(\boldsymbol{x})$ obtained by evolving in time a fixed initial state described by the density matrix $\hat{\rho}_0$ [see Fig. 2 for a specific example of the encoding process described by the evolution map $\mathcal{M}(\boldsymbol{x})$]. The measured features are represented by a vector of observables $\boldsymbol{\phi}(\boldsymbol{x})$ (with an added 1 corresponding to the unity operator as the first element to create an offset term) that belongs to the feature space $\mathcal{F}$. The trial function is obtained by applying a linear transformation to the feature vector (depending nonlinearly on $\boldsymbol{x}$) with a vector of weights $\boldsymbol{w}$ that is optimized via the training procedure described in the main text.

where

$$\langle \hat{O}_j \rangle_{\boldsymbol{x}} = \text{Tr}[\hat{O}_j \hat{\rho}(\boldsymbol{x})]. \tag{6}$$

Here we consider ensemble measurements and hence the order of the observables in $\mathcal{O}$ is irrelevant. The vector $\boldsymbol{\phi}(\boldsymbol{x})$ belongs to the feature space $\mathcal{F} \subseteq \mathbb{R}^P$ and depends on the input $\boldsymbol{x}$, generally in a nonlinear fashion. Note that the constant component 1, which ensures that the trial function can fit a biased target function, can be seen as the measurement of the identity observable, since the density matrix $\hat{\rho}(\boldsymbol{x})$ always has unit trace. Finally, the trial function f of the noisy quantum kernel machine, which depends on the vector of variational parameters $\boldsymbol{w}$, is given by the affine transformation:

$$f : \boldsymbol{x} \mapsto \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \tag{7}$$

where the vector $\boldsymbol{w} \equiv (b, w_1, \ldots, w_P)^T \in \mathbb{R}^{P+1}$ contains the parameters of the linear transformation and b represents the bias term. An alternative approach to the construction of the feature vector, based on time multiplexing measurements, will be presented in Sec. IV.

C. Training procedure

A trial function characterized by its weights $\boldsymbol{w}$ can be optimized using a regularized least-squares loss function over a training set $(\boldsymbol{x}_i, y_i) \in \mathcal{S}$ consisting of N_{train} inputs $\boldsymbol{x}_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$, namely,

$$\mathcal{L}(\boldsymbol{w}|\mathcal{S}) := \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} [y_i - \boldsymbol{w}^T \boldsymbol{\phi}(x_i)]^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2. \tag{8}$$

The second term in Eq. (8) is a regularization penalty that helps to prevent overfitting. The corresponding regularization parameter λ controls the strength of the overfitting penalty. Adding such a regularization bias is on average equivalent to adding a centered Gaussian noise of variance λ to the measurement features before the optimization [2]. Such classifiers are known as least-square support-vector classifiers [35]. Al-

though most classification problems are commonly treated with other loss functions [36], using the least-squares loss function allows us to perform the optimization analytically. Indeed, upon introducing the $(P + 1) \times N_{\text{train}}$ matrix $\boldsymbol{\Phi}$, the columns of which are the quantum feature vectors $\boldsymbol{\phi}(\boldsymbol{x}_i)$, associated to the training input $\boldsymbol{x}_i$, and $\boldsymbol{y}$, the column vector of size N_{train} containing the corresponding labels, the optimal weights are given by [37]

$$\boldsymbol{w}^* = (\boldsymbol{\Phi}\boldsymbol{\Phi}^T + N_{\text{train}}\lambda\mathbb{1})^{-1}\boldsymbol{\Phi}\boldsymbol{y}. \tag{9}$$

III. QUANTUM KERNEL AND DECOHERENCE

The generic encoding-decoding scheme encompasses a large class of quantum machine-learning models. Here, we describe a decoding based on a linear combination of measurements, but other decoding methods were proposed in the literature. In particular, it was recently shown that quantum neural networks can be mapped to models with an encoding-decoding structure [38], where the decoding is achieved by optimizing a single parametrized measurement. Models described by the previous scheme can be analyzed in the framework of kernel theory, which provides useful tools to understand properties such as expressivity, trainability, and capacity to generalize to a test sample of unseen data. In this section we first concisely introduce the kernel framework. We then specialize our discussion to noisy quantum kernels, and show how we can link the role of dissipation and decoherence to the kernel's main figures of merit. We aim at determining the largest class of functions that can be approximated by our trial function f. This class depends on the type of decoding used, that is, on the specific set of measurements that are performed on the quantum system. When measuring a set $\mathcal{O}$ of observables, this function space reads

$$\mathcal{H}(\mathcal{O}) = \{f : \boldsymbol{x} \mapsto \text{Tr}[\hat{\rho}(\boldsymbol{x})\hat{A}] \mid \hat{A} \in \text{Span}(\mathcal{O})\}. \tag{10}$$

In this case, the feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ gives rise to a positive semidefinite and symmetric function which we call the feature

kernel:

$$k_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}'). \tag{11}$$

This kernel function, together with the probability distribution p of inputs $\boldsymbol{x} \in \mathcal{X}$[39], uniquely determines a specific set of real-valued functions, the so-called reproducing kernel Hilbert space (RKHS):

$$\text{Span}\{f : \boldsymbol{x} \mapsto k_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{x}') \mid \boldsymbol{x}' \in \mathcal{X}\}. \tag{12}$$

The RKHS associated to $k_{\mathcal{O}}$ can be shown to be exactly the space of hypothesis functions $\mathcal{H}(\mathcal{O})$ [40]. Hence, the study of the kernel function allows one to investigate the structure of $\mathcal{H}(\mathcal{O})$. In particular, it follows that one can use the eigendecomposition of the kernel function as a basis of the class of functions that can be represented by our model. This useful property motivates the adoption of a kernel standpoint in what follows.

A. Quantum kernel

In order to discuss the expressive power of our model, we introduce the largest class of transformations $\mathcal{H}_{\text{full}}$ that can be achieved for a given encoding strategy [41]:

$$\mathcal{H}_{\text{full}} = \{f : \boldsymbol{x} \mapsto \text{Tr}[\hat{\rho}(\boldsymbol{x})\hat{A}] \mid \hat{A} = \hat{A}^\dagger\}. \tag{13}$$

The class of transformation yielded by a set of measurements $\mathcal{O}$ is necessarily included in this maximal class $\mathcal{H}(\mathcal{O}) \subseteq \mathcal{H}_{\text{full}}$; the equality holds whenever $\mathcal{O}$ is a complete set of observables. In the following, we will use the term "full tomography" to refer to this ideal implementation. It turns out that $\mathcal{H}_{\text{full}}$ is the RKHS of a particular kernel, the quantum kernel, that solely depends on the feature map $\hat{\rho}(\boldsymbol{x})$ [22,42]:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \text{Tr}[\hat{\rho}(\boldsymbol{x})\hat{\rho}(\boldsymbol{x}')]. \tag{14}$$

This kernel is simply the Hilbert-Schmidt inner product between quantum states encoding inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$. As it represents the maximal achievable class of transformation an encoding can give, the quantum kernel provides insight on the expressive power of our model. Note that this kernel can be identified with the previous feature kernel $k_{\mathcal{O}}$ provided that the measurements $\mathcal{O}$ form an orthonormal basis (with respect to the Hilbert-Schmidt inner product) $\mathcal{B} = \{B_j\}_j$ of the space of observables, i.e., $k = k_{\mathcal{B}}$ with $\text{Tr}[\hat{B}_i\hat{B}_j] = \delta_{ij}$, and we impose $\hat{B}_0 \propto \hat{\mathbb{1}}$ by convention. In what follows, it will be useful to work with a "centered" version of the quantum kernel. Centering the kernel is equivalent to working with hypothesis functions that have zero mean value on the input set. As we will show, this is convenient for interpreting some of the key quantities we will introduce in terms of probabilistic quantities. In Appendix B, we show that, at least for balanced data, the use of the L2 loss function allows us to work with a centered version of the quantum kernel without lack of generality. The centered kernel is given by

$$k_c(\boldsymbol{x}, \boldsymbol{x}') = \text{Tr}[\delta\hat{\rho}(\boldsymbol{x})\delta\hat{\rho}(\boldsymbol{x}')] \tag{15}$$

where as before we define

$$\delta\hat{\rho}(\boldsymbol{x}) = \hat{\rho}(\boldsymbol{x}) - \mathbb{E}_p[\hat{\rho}(\boldsymbol{x})] \tag{16}$$

and the corresponding RKHS is

$$\mathcal{H}_{k,c} = \text{Span}\{f : \boldsymbol{x} \mapsto k_c(\boldsymbol{x}, \boldsymbol{x}') \mid \boldsymbol{x}' \in \mathcal{X}\}. \tag{17}$$

The constant feature we introduced in Eq. (5) becomes irrelevant when using centered quantities, so we drop it and define

$$\delta\boldsymbol{\phi}(\boldsymbol{x}) \equiv (\delta\langle\hat{O}_1\rangle_{\boldsymbol{x}}, \ldots, \delta\langle\hat{O}_P\rangle_{\boldsymbol{x}})^T. \tag{18}$$

We can also correspondingly drop the weight term b, so that the weight vectors can be redefined as $\boldsymbol{w} = (w_1, \ldots, w_P)^T \in \mathbb{R}^P$. The space $\mathcal{H}_{k,c}$ can be rewritten as

$$\mathcal{H}_{k,c} = \{f : \boldsymbol{x} \mapsto \boldsymbol{w}^T\delta\boldsymbol{\phi}(\boldsymbol{x}), \quad \boldsymbol{w} \in \mathbb{R}^P\}, \tag{19}$$

where the centered quantum kernel reads

$$k_c(\boldsymbol{x}, \boldsymbol{x}') = \delta\boldsymbol{\phi}(\boldsymbol{x})^T\delta\boldsymbol{\phi}(\boldsymbol{x}') \tag{20}$$

with the choice of $\mathcal{O} = \mathcal{B}$. The quantum feature matrix $\boldsymbol{\Phi}$ is then replaced by a $P \times N_{\text{train}}$ matrix $\delta\boldsymbol{\Phi}$, the columns of which are the centered feature vectors $\delta\boldsymbol{\phi}(\boldsymbol{x}_i)$.

B. Kernel eigendecomposition

Under general assumptions, the centered quantum kernel admits a decomposition into an orthonormal family of eigenfunctions [40]:

$$k_c(\boldsymbol{x}, \boldsymbol{x}') = \sum_i \lambda_i\delta\psi_i(\boldsymbol{x})\delta\psi_i(\boldsymbol{x}'), \quad \mathbb{E}_p[\delta\psi_i\delta\psi_j] = \delta_{ij}, \tag{21}$$

where $\{\lambda_i\}_i$ are positive eigenvalues sorted in a decreasing order, namely, $\lambda_{i+1} \leqslant \lambda_i, \forall i$. When necessary, we can complete this orthonormal family into a basis with eigenfunctions associated to zero eigenvalues. In the case of the uncentered quantum kernel, the kernel eigenfunctions correspond to an orthonormal basis of system observables [26]. When the kernel is centered, the basis of kernel eigenfunctions corresponds to an orthonormal basis $\{\hat{E}_i\}_i$ of the space of zero-trace observables, which we call eigenobservables. Such operators satisfy the following properties:

$$\text{Tr}[\hat{E}_i\hat{E}_j] = \delta_{ij}, \quad \text{Tr}[\hat{E}_i] = 0. \tag{22}$$

The eigenfunctions are given by

$$\delta\psi_i(\boldsymbol{x}) = \frac{1}{\sqrt{\lambda_i}}\text{Tr}[\delta\hat{\rho}(\boldsymbol{x})\hat{E}_i] = \frac{1}{\sqrt{\lambda_i}}\delta\langle\hat{E}_i\rangle_{\boldsymbol{x}}. \tag{23}$$

The corresponding eigenvalues are then given by the variances of the eigenobservable measurements over the input set, namely,

$$\lambda_i = \mathbb{E}_p[\delta\langle\hat{E}_i\rangle_{\boldsymbol{x}}^2] = \text{Var}_p[\langle\hat{E}_i\rangle_{\boldsymbol{x}}]. \tag{24}$$

One can see this eigendecomposition of the kernel as a principal-component analysis in the space of quantum features, as it yields an orthogonal basis of measurement functions ordered by their variances on the input set. We stress that these are variances of the observables' expectation values over the quantum states representing the different inputs, and thus are very different from the quantum variance of the corresponding observable for a specific state. The previous decomposition of the kernel is very useful for grasping the learning mechanism and the model expressivity. Upon working with centered features, the loss function introduced in

Eq. (8) becomes

$$\mathcal{L}_c(\boldsymbol{w} \mid \mathcal{S}) = \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} [y_i - \boldsymbol{w}^T \delta\boldsymbol{\phi}(x_i)]^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2. \quad (25)$$

Following [36], we can decompose the trial function $f(\boldsymbol{x}) = \boldsymbol{w}^T \delta\boldsymbol{\phi}(\boldsymbol{x})$ in the basis of the kernel eigenfunctions, namely, as $f(\boldsymbol{x}) = \sum_j \beta_j \delta\psi_j(\boldsymbol{x})$. Exploiting such decomposition, the loss function becomes

$$\mathcal{L}_c(\boldsymbol{\beta} \mid \mathcal{S}) = \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left[y_i - \sum_j \beta_j \delta\psi_j(\boldsymbol{x}_i) \right]^2 + \frac{\lambda}{2} \sum_j \frac{\beta_j^2}{\lambda_j}. \quad (26)$$

Note that in the regularization term the components of the trial function on the eigenbasis are weighted by the corresponding kernel eigenvalues. The lower the variance of an eigenobservable, the more the corresponding eigenfunction is penalized. Hence the regularization parameter λ acts as a smooth cutoff on the basis of the kernel eigenfunctions, which are then used to approximate the target function. The spectrum of the kernel characterizes the generalization capacity and the expressivity of our model. It also finds applications in understanding many other machine-learning scenarios. For instance, in the context of classical neural networks it has links with learning curves [43,44]. Moreover, the kernel (or the neural tangent kernel in the context of classical and quantum neural networks [11,45,46]) shares its spectrum with the Fisher information matrix, of particular relevance for quantum neural networks [47].

C. Role of decoherence on expressivity and generalization error

The exponential growth of the Hilbert-space dimension with the number of qubits in a network and the complex dynamics of quantum systems have created hope for a quantum advantage in the field of quantum machine learning. However, it is known that having a very high-dimensional feature space does not necessarily guarantee high machine-learning performances [36,48]. Indeed, recent investigations within the quantum kernel framework somehow mitigated the hope for a general quantum advantage [26,28,49]. Yet, a clear quantum advantage has been demonstrated for some specific tasks [25,46], again by exploiting the quantum-kernel formalism. In order for a quantum-kernel-based model to perform well on a given task, the set of transformations achieved must be well "aligned" with the target function $y(\boldsymbol{x})$. This notion of alignment is mathematically encapsulated in the kernel-target-alignment measure [50], which reads, for the centered quantum kernel,

$$A(k_c, y) = \frac{\mathbb{E}_p[y(\boldsymbol{x}) k_c(\boldsymbol{x}, \boldsymbol{x}') y(\boldsymbol{x}')]}{\mathbb{E}_p[k_c(\boldsymbol{x}, \boldsymbol{x}')^2]^{1/2} \mathbb{E}_p[y(\boldsymbol{x})^2]}$$

$$= \frac{\sum_i \lambda_i \mathbb{E}_p[\delta\psi_i(\boldsymbol{x}) y(\boldsymbol{x})]^2}{\left(\sum_i \lambda_i^2\right)^{1/2} \mathbb{E}_p[y(\boldsymbol{x})^2]}. \quad (27)$$

Although the kernel-target alignment measures how well a kernel and the associated embedding fit a specific function, in this paper we introduce another figure of merit that does not depend on a specific task, namely, the "effective kernel rank"

$R_{\text{eff}}(k)$, which quantifies the effective number of independent transformations that a given kernel can yield. Such a quantity is defined as

$$\sqrt{R_{\text{eff}}(k_c)} = \sum_j A(k_c, g_j), \quad (28)$$

where $\{g_j\}_j$ is any orthonormal basis of functions on the input space. As shown in Appendix A1, for the centered quantum kernel, the effective kernel rank can be also expressed in terms of variances of the quantum expectation values of the measured observables:

$$\sqrt{R_{\text{eff}}(k_c)} = \frac{\sum_{i=1}^P \text{Var}_p[\langle \hat{O}_i \rangle_{\boldsymbol{x}}]}{\left(\sum_{i,j=1}^P \text{Cov}_p[\langle \hat{O}_i \rangle_{\boldsymbol{x}}, \langle \hat{O}_j \rangle_{\boldsymbol{x}}]^2 \right)^{\frac{1}{2}}}. \quad (29)$$

Note that the denominator acts as a normalization and can be seen as a measure of the redundancy of the embedding when expressed in terms of $\hat{O}_i$. In Sec. V, we will investigate in a rather general class of physical models how the kernel effective rank scales with the system size and with noise. In Appendix A1, we also provide the proof showing that the kernel effective rank can be expressed in terms of the kernel spectrum:

$$\sqrt{R_{\text{eff}}(k)} = \frac{\sum_i \lambda_i}{\sqrt{\sum_i \lambda_i^2}}. \quad (30)$$

This expression is reminiscent of the reciprocal of the inverse participation ratio. The kernel effective rank provides information about the size of its support. Moreover, we have the following inequality:

$$R_{\text{eff}}(k) \leqslant |\{\lambda_i \neq 0\}|. \quad (31)$$

This is saturated when all the nonzero eigenvalues are equal. The numerator in the expression for the square root of the effective kernel rank is the kernel trace, which can be rewritten as

$$\sum_i \lambda_i = \mathbb{E}_p[\text{Tr}[\hat{\rho}(\boldsymbol{x})^2]] - \text{Tr}[\mathbb{E}_p[\hat{\rho}(\boldsymbol{x})]^2]. \quad (32)$$

In this expression, we recognize the difference between the average purity of the embedded density matrices over the input space and the purity of the average embedding matrix. The first term is of great relevance to our analysis, as it crucially depends on the dissipation and decoherence affecting the noisy quantum system: Indeed, a low purity is the consequence of the openness of the quantum system. The second term instead measures the diversity of the embedding map; its importance is discussed in [26]. We emphasize that the kernel trace also appears to be relevant when investigating the ability of the model to perform well on unseen data, and hence on its generalization properties. To measure the performance of a model on a binary classification task we use the accuracy $\mathcal{A}$. Given a prediction function f, the accuracy is given by the fraction of samples for which f assigns the right label and it can be defined as the expectation of a 0-1 loss function:

$$\mathcal{A}(f) = \mathbb{E}[\mathbb{1}_{y(\boldsymbol{x}) f(\boldsymbol{x}) \geqslant 0}]. \quad (33)$$

Since during the training we only have access to the data set $\mathcal{S}$ and not to the true distribution p, expectation values

can only be approximated using the empirical distribution $\hat{p}$ on $\mathcal{S}$. The corresponding empirical expectations are given by $\mathbb{E}_{\hat{p}}[f(\boldsymbol{x})] = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} f(\boldsymbol{x}_i)$. From this, we can define the empirical accuracy $\mathcal{A}$ on the training set $\mathcal{S}$ and the true accuracy $\mathcal{A}^*$. Correspondingly, we can introduce the risk $\mathcal{R}^*$, also called error or inaccuracy, as $\mathcal{R}^* = 1 - \mathcal{A}^*$ (its empirical counterpart is defined analogously). It is convenient to introduce slightly modified versions of the risk and inaccuracy that depend on a margin parameter $\eta > 0$. We introduce the η-margin loss as

$$\Phi_\eta(y) = \begin{cases} 1 & \text{if} \quad y \leqslant 0 \\ 1 - \frac{y}{\eta} & \text{if} \quad 0 \leqslant y \leqslant \eta \, . \\ 0 & \text{if} \quad \eta \leqslant y \end{cases} \tag{34}$$

Correspondingly, we can introduce the empirical η-margin risk as

$$\mathcal{R}_\eta(f) = \mathbb{E}_{\hat{p}}[\Phi_\eta[y(\boldsymbol{x})f(\boldsymbol{x})]]. \tag{35}$$

The η-margin risk and the risk satisfy the following inequality:

$$\mathcal{R}(f) \leqslant \mathcal{R}_\eta(f) \leqslant \mathbb{E}_{\hat{p}}[\mathbb{1}_{y(\boldsymbol{x})f(\boldsymbol{x})\leqslant\eta}]. \tag{36}$$

The ability of the model to generalize well on unseen data is then quantified by the generalization error:

$$\mathcal{E} = \mathcal{R}^* - \mathcal{R}. \tag{37}$$

For kernel methods with kernel k, the generalization error admits an upper bound involving the $N_{\text{train}} \times N_{\text{train}}$ empirical kernel matrix $\boldsymbol{K}$ the entries of which are defined as $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This bound depends on the specific task under consideration and on the exact space of trial functions used (details on the bound used and its derivation can be found in [51] and in Appendix A2). To derive the upper bound, we fix a class of trial functions of the form $f : \boldsymbol{x} \mapsto \boldsymbol{w}^T \delta\boldsymbol{\phi}(\boldsymbol{x})$ where $\delta\boldsymbol{\phi}(\boldsymbol{x})$ corresponds to measurements of an orthonormal basis of the observable: $\delta\phi_i(\boldsymbol{x}) = \delta\langle\hat{B}_i\rangle_{\boldsymbol{x}}$. We further constrain this class by choosing a parameter $\Lambda \geqslant 0$, and require that the trial function's parameters $\boldsymbol{w}$ satisfy $\|\boldsymbol{w}\|^2 \Lambda \leqslant 1$. By exploiting Eq. (32), we get that, for such functions, the following inequality holds with probability at least $1 - \delta$ on the training set $\mathcal{S}$:

$$\mathcal{R}^*(f) - \mathcal{R}_\eta(f) \leqslant \frac{2}{\eta} \left(\frac{\mathbb{E}_{\hat{p}}[\text{Tr}[\hat{\rho}^2]] - \text{Tr}[\mathbb{E}_{\hat{p}}[\hat{\rho}]^2]}{N_{\text{train}} \Lambda} \right)^{\frac{1}{2}} + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2N_{\text{train}}}}. \tag{38}$$

Other generalization bounds can be established; in particular the authors of [52] found another bound using a quantum information theory standpoint, and their conclusions are in agreement with our results. Let us make a few important comments on the meaning of this inequality. The inequality has a probabilistic character controlled by $\delta > 0$. If we set this parameter to 0^+, the bound is always satisfied although it becomes trivial. The same goes with the margin parameter η: as $\eta \to 0^+$ the margin error $\mathcal{R}_\eta(f)$ tends to the training error $\mathcal{R}(f)$, but again the right-hand side of the inequality diverges. The parameter Λ is another sort of regularization parameter, as the parameter λ: If $\Lambda \to 0^+$, the norm of the weight vector $\|\boldsymbol{w}\|$ can be arbitrarily large and overfitting is not limited.

Correspondingly, the right-hand side diverges and the bound becomes trivial. The most important crucial physical quantity involved in the upper bound is the kernel trace given in Eq. (32). Such a quantity accounts for the model expressivity. This duality between expressivity and generalization is crucial in machine learning [36]. What is relevant to our analysis is that this expressivity measure involves the mean purity of the embedded states and hence is affected by dissipation and decoherence acting on the noisy quantum kernel machine. The appearance of the regularization parameter Λ in this upper bound is also relevant as it allows us to establish a link with experimental constraints, such as imperfect measurements. In fact, as we will see in Sec. V, adding a Gaussian error of standard deviation σ to the observable measurements is equivalent to working with an infinitely precise measurement apparatus while replacing the regularization parameter λ with $\lambda + \sigma^2$ [2].

IV. NOISY QUANTUM KERNEL MACHINES WITH DRIVEN-DISSIPATIVE SPIN CHAINS

As an illustrative example, we here numerically simulate noisy quantum kernel machines based on one-dimensional chains of spins subject to both driving and decoherence. The simulation of such an open quantum system for a large number of inputs, various choices of the number of sites, and distinct disorder realizations is a computationally daunting task [53]. Indeed, this requires us to exactly integrate a large set of corresponding Lindblad master equations of the form of Eq. (2). Hence, we have considered a simplified classification task involving only a subset of the Modified National Institute of Standards and Technology (MNIST) dataset, namely, classifying images of handwritten digits corresponding to the digits 3, 6, and 8, which share common shapes. A schematic description of the task and of the feature encoding through driving of the considered physical system is presented in Fig. 2. The original MNIST dataset consists of 28×28-pixel images. Encoding such high-dimensional features in the state of a quantum system is not an easy task. Therefore, we first linearly downsample the raw images from 28×28 to 8×8 pixels, thereby reducing the dimension of the input features. The downsampled images, viewed as vectors, are then multiplied by a random $8^2 \times 10$ matrix $\boldsymbol{W}$, the entries of which are uniformly drawn over the interval $[-1, 1]$, yielding vectors $\boldsymbol{x}' = (x_1', \ldots, x_M')^T$ of $M = 10$ random-projection features. These are finally normalized by three times the standard deviation of the set $\{x_i' \mid i = 1, \ldots, M, \ x \in \mathcal{S}\}$. At the end of this procedure, every image in the dataset is represented by a vector $\boldsymbol{x}$ of size $M = 10$, which will be used as inputs in the following. These are computed only once and reused throughout this paper, except in Sec. VB.

This encoding is designed so as to fix the amount of information fed to the system, independently from its number of sites. It allows us to perform a fair comparison of models associated to quantum systems of increasing sizes. In particular, this ensures that any observed increase of the performance with the system size is solely due to an intrinsic enhancement of the model expressive power. In Sec. VB, we lift the above-defined "information bottleneck" and use
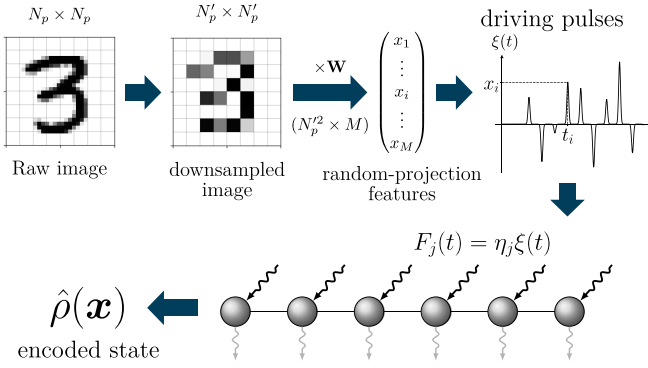
FIG. 2. Schematic representation of the encoding procedure for the MNIST classification task. The input grayscale image, of original size $N_p \times N_p$, with $N_p = 28$, is first downsampled to a size of $N_p' \times N_p'$ (or $N_p'^2 \times 1$ when viewed as a column vector), with $N_p' = 8$, and linearly transformed by a fixed $N_p'^2 \times M$ random projection filter $\mathbf{W}$ to yield the vectors $\mathbf{x}'$ containing $M = 10$ random-projection features. Those features are normalized by three times the standard deviation over the set of all features for all images in the training set, and we denote $\mathbf{x}$ the normalized vectors representing the images. The vector $\mathbf{x}$ is then encoded into a sequence of driving pulses $\xi(t)$, where the amplitude of the ith pulse (at time t_i) is proportional to the input's ith component x_i. Finally, the pulses are used to drive a spin chain (initially prepared in the state $\hat{\rho}_0$), where the driving amplitude at site j is $F_j(t) = \eta_j \xi(t)$ with η_j a random site-dependent scale factor. We define the state of the spin chain immediately after the driving sequence to be the encoded state, represented by its density matrix $\hat{\rho}(\mathbf{x})$.

a different encoding, where the number of encoded features M scales with the system size N. Therein, we show that this results in competitive performances, as compared to classical reservoir-computing settings involving hundreds to thousands of degrees of freedom [6,8]. In what follows, we denote $\mathcal{X} \subseteq \mathbb{R}^M$ the input space consisting of the random-projection features representing the images to classify, and $\mathcal{Y} = \{3, 6, 8\}$ the set of corresponding labels. Our dataset consists of 17 000 images, which we split into a training set of $N_{\text{train}} = 15\,000$ images and a testing set of $N_{\text{test}} = 2000$ images. As before, the training set is denoted as $\mathcal{S} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, N_{\text{train}}\}$. The system in which we encode the previous features is a driven-dissipative one-dimensional chain of N spins 1/2 described by the following Heisenberg XYZ Hamiltonian:

$$\hat{H}(t;\mathbf{x}) = \frac{\hbar}{2} \sum_{i=1}^{N} \left[F_i(t;\mathbf{x}) \hat{\sigma}_x^i + \Delta_i \hat{\sigma}_z^i \right]$$
$$- \frac{\hbar}{2} \sum_{\langle i,j \rangle} \left(J_{ij}^x \hat{\sigma}_x^i \hat{\sigma}_x^j + J_{ij}^y \hat{\sigma}_y^i \hat{\sigma}_y^j + J_{ij}^z \hat{\sigma}_z^i \hat{\sigma}_z^j \right), \quad (39)$$

with $F_i(t;\mathbf{x})$ an input-dependent driving field, Δ_i an on-site frequency detuning, and J_{ij}^k the symmetric coupling rate between nearest neighbors. Here, indices $\langle i, j \rangle$ run over all pairs of nearest neighbors. Parameters J_{ji}^k and Δ_i are uniformly drawn at random in the interval $[0, 2J]$. $1/J$ will be used as the unit of time in the numerical plots. We prepare the system in an initial state with all spins down $\hat{\rho}_0 = $
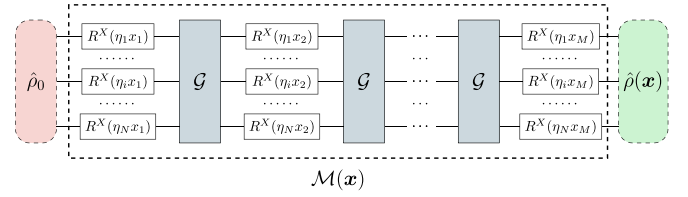

FIG. 3. Equivalent circuit of the encoding procedure for the MNIST classification task. If the driving pulses are sharp enough, the encoding process of Fig. 2 can be equivalently seen as a quantum circuit, where the ith driving pulse on site j is effectively a single-qubit X-rotation gate $R^X(\eta_j x_i)$, and the pulses at different times are separated by the gate $\mathcal{G}$ generated by the free dynamics of the spin system in the absence of the drive. Note that the entire process between $\hat{\rho}_0$ and $\hat{\rho}(x)$ serves as the dynamical map $\mathcal{M}(x)$ shown in Fig. 1.

$\bigotimes_{i=1}^{N} |0\rangle\langle 0|$. The encoding of the input $\mathbf{x}$ corresponding to a given image into the system state is performed by driving the system with a series of $M = 10$ sharp Gaussian pulses, the amplitudes of which are proportional to the input vector elements, as illustrated in Fig. 2. We first define a generic driving $\xi(t;\mathbf{x})$ from the feature $\mathbf{x}$:

$$\xi(t;\mathbf{x}) = \sum_{k=1}^{N} \frac{x_k}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t - t_k)^2}{2\sigma^2} \right),$$
$$t_k = (k-1)\Delta t + 10\sigma, \quad \forall k = 1, \ldots, M, \quad (40)$$

where the time interval between two successive pulses is $\Delta t = 1/(2J)$ and the width of each pulse is $\sigma = 1/(50J)$. Then the driving on site i is taken to be proportional to this generic driving:

$$F_i(t;\mathbf{x}) = \eta_i \xi(t;\mathbf{x}), \quad (41)$$

where the η_i are random factors uniformly distributed in the interval $[-\pi, \pi]$. Under these driving conditions, the coherent part of the system dynamics can be thought of as that of an equivalent quantum circuit alternating between a set of local X-rotation gates, of the form $R_i^X(\eta_i x_k)$, and a deep block generating entanglement among qubits [54], as illustrated in Fig. 3. The scaling factors η_i prevent the spins from rotating all together. This procedure, where a random-projection feature is fed to the system every Δt, is in close analogy with the repeated-encoding prescription in variational quantum circuits, which is known to improve the expressivity of a model [41]. Shortly after the last pulse of the driving ends, at time $\tau = 30\sigma + M\Delta t$, we get the final encoded state represented by the density matrix $\hat{\rho}(\mathbf{x})$. This encoding procedure acts as a nonlinear map from the input space of images to the high-dimensional space of N-spin mixed quantum states. Concerning the nonunitary dynamics due to the openness of the quantum kernel machine, we will consider spin dephasing as the source of decoherence. Within the Lindblad master equation formalism [Eq. (2)], this process is described by the jump operators $\hat{A}^j = \hat{\sigma}_z^j$, and we consider a uniform dephasing rate for each site $\gamma_i = \gamma, \forall i \in \{1, \ldots, N\}$. Note that while the considered illustrative task involves three classes, it can be reduced to a set of binary classification problems by changing the labels $y \in \{3, 6, 8\}$ into vector labels of the form

$(y_1, y_2, y_3)^T$ with $y_j \in \{-1, 1\}^3$. For example, an outcome $(-0.3, -0.2, 0.9)$ would correspond to the digit 8. This "one-vs-the-rest" approach is equivalent to training three binary classifiers, one for each class, and takes the highest output among the three classifiers as a prediction. With this approach the training effort scales linearly with the number of classes. However, for the sake of simplicity, we will use binary classification notations in the following, and consider that the labels belong to $\{-1, 1\}$. Regarding the measurements of the system observables, we will consider two measurement protocols. (i) A *full tomography* of the output density matrix: In this case we consider that the measurements are made without delay after the end of the encoding, and the extracted features are exactly the components of the generalized Bloch vector $\boldsymbol{\phi}(\boldsymbol{x})$ by considering a complete set of observables. (ii) A *time multiplexing* measurement protocol: In this case the output is obtained by sequential measurements at different times of a set of local observables.

A. Full tomography

Any Hermitian operator of the considered spin system can be decomposed on the orthogonal (for the Hilbert-Schmidt inner product) basis of Pauli strings. For a system of N spins, we write this basis $\{\hat{O}_i \,|\, i = 0, \ldots, P\}$, with $P = 4^N - 1$. The corresponding observables are such that

$$\hat{O}_i = \bigotimes_{k=1}^{N} \hat{\sigma}_{i_k}^k, \;\; i_k \in \{0, 1, 2, 3\}, \;\; \mathrm{Tr}[\hat{O}_i^\dagger \hat{O}_j] = 2^N \delta_{ij}, \;\; \forall i, j, \tag{42}$$

with $\hat{O}_0 = \hat{\mathbb{1}}$, and thus any observable $\hat{A}$ is decomposed in this basis through the expansion:

$$\hat{A} = \frac{1}{2^N} \left(\mathrm{Tr}[\hat{A}] \hat{\mathbb{1}} + \sum_{i=1}^{P} \mathrm{Tr}[\hat{O}_i \hat{A}] \hat{O}_i \right). \tag{43}$$

The density matrix associated to the input $\boldsymbol{x}$ can also be decomposed into this basis:

$$\hat{\rho}(\boldsymbol{x}) = \frac{1}{2^N} \left(\hat{\mathbb{1}} + \sum_{i=1}^{P} \langle \hat{O}_i \rangle_{\boldsymbol{x}} \hat{O}_i \right), \tag{44}$$

and hence any density matrix is uniquely characterized by its associated generalized Bloch vector. For the full-tomography decoding we take these Bloch vectors as the quantum features, which is equivalent to rescaling the quantum kernel function [Eq. (14)] by a constant factor of 2^N. The encoding method we use leads to embedded states that exhibit entanglement. Figure 4(a) shows that the average entanglement negativity quickly increases during the encoding, and then eventually decays at a rate depending on γ. In parallel, as we see from Fig. 4(b), there is a finite von Neumann entropy of the system due to mixed character of the state. In Sec. V, we will show how these processes affect the performances of noisy quantum kernel machines.

B. Time multiplexing measurements

A simplified and experimentally less demanding decoding is obtained by measuring all the single-site observables
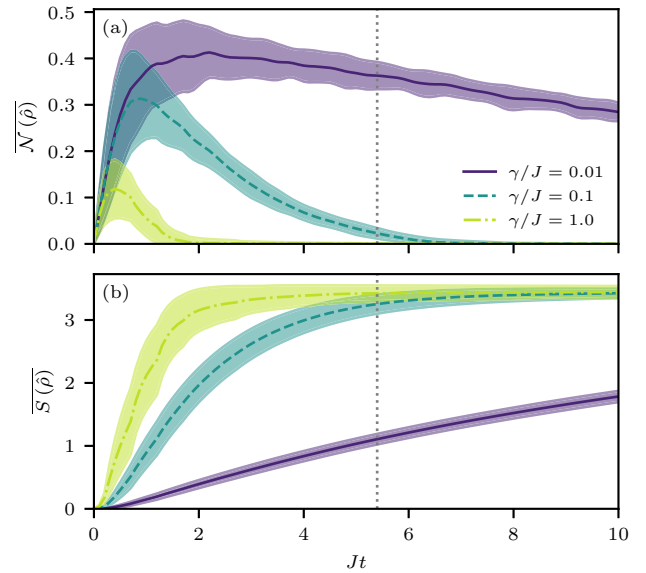

FIG. 4. Time dynamics of the average entanglement negativity $\overline{\mathcal{N}(\hat{\rho})}$ (a) and the average von Neumann entropy $\overline{S(\hat{\rho})}$ (b) in presence of pure dephasing with different values of the corresponding rates γ. At the initial time $t = 0$ the system is in the pure state $\hat{\rho}_0$ defined in the text. Note that the driving sequence finishes at the time $Jt \simeq 5$ indicated by the vertical dotted lines. The time is expressed in units of $1/J$ where J is the average value of the spin coupling. We define $\overline{\mathcal{N}(\hat{\rho})}$ as the average over all the sites of the negativities associated to the system partitions having the form $\{\{\text{site } i\}, \{\text{site } j \,|\, j \neq i\}\}$. This quantity is averaged over 20 inputs $\boldsymbol{x} \in \mathcal{X}$, over five disordered configurations of spin couplings, and for a chain of $N = 5$ spins. The filled areas correspond to a one standard deviation confidence interval.

(i.e., the three local Pauli spin operators) at different times after the end of the encoding. In the following, we will denote N_{rep} the number of repetitions of these measurements. Hence, for a system of N spins, a total number $3N \times N_{\text{rep}}$ of measurements have been performed after N_{rep} repetitions. We use measurements of the on-site observables for each spin, which correspond to the components of the Bloch vectors of the reduced density matrices on each site. We consider corresponding observables in the Heisenberg picture. The new feature vector $\tilde{\boldsymbol{\phi}}(\boldsymbol{x})$ in the time multiplexing protocol has entries of the form $\langle B_i(t + k\delta t_m) \rangle_{\boldsymbol{x}}$ with $1 \leqslant i \leqslant 3N$, $1 \leqslant k \leqslant N_{\text{rep}}$, where δt_m is the time interval between two consecutive measurements. Similar methods were used in previous works to perform an approximate tomography of the system state [55,56]. Note that the time multiplexing procedure can only decrease the model expressive power when compared to the full tomography, as information leaks into the system's environment as the system evolves between successive measurement times (see Appendix C).

V. NUMERICAL RESULTS

In this section, we discuss the numerical results on the noisy quantum kernel machines obtained by considering the model spin Hamiltonian, dephasing channels, input encoding
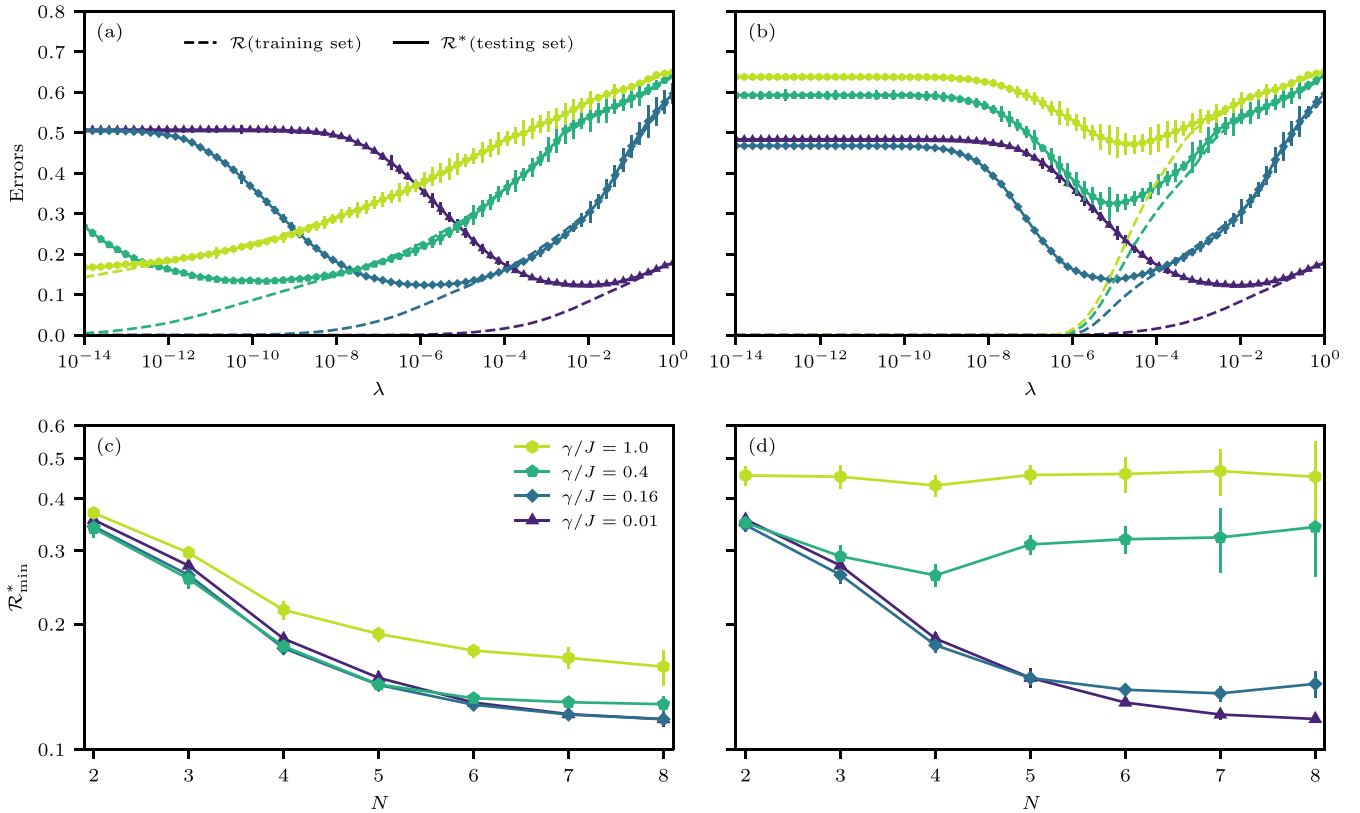
FIG. 5. (a) Training error $\mathcal{R}$ (dashed lines) and testing error $\mathcal{R}^*$ (solid lines and markers) as a function of the regularization parameter λ for a chain with $N = 7$ spins in the presence of pure dephasing for different values of the corresponding rate γ (different markers in the legend) in units of the average spin coupling J. Measurements are assumed to be ideal. (b) Same as panel (a) with an extra random Gaussian noise of width $\sigma = 10^{-3}$ added to the observable expectation values to account for imperfect measurements. For each value of λ the corresponding errors are averaged over 15 disordered configurations, and the error bars are bootstrap estimates of the standard deviation for the estimated mean values. We use ten bootstrap sets, each consisting of 15 samples randomly drawn with replacement from the original set of 15 disorder realizations. (c) Minimal testing error as a function of the number of spins N for different values of the dephasing rate γ. For each disorder configuration, the regularization parameter λ is chosen to minimize the testing error and the resulting minimum is averaged over the disorder. The error bars are derived using the same bootstrap procedure. Number of disorder configurations: 50 for $N = 2$ to 5 spins, 25 for $N = 6$, 15 for $N = 7$, and 5 for $N = 8$. (d) Same as panel (c) with an extra random Gaussian noise of width $\sigma = 10^{-3}$ added to the observable expectation values to account for imperfect measurements.

via driving, decoding protocol through measurement, and the classification task detailed in the previous section.

A. Performances, noise, and system size

The main goal is to determine how the performance of the noisy quantum kernel machine scales with the amount of noise and the number of chain sites, i.e., network nodes. To provide a fair comparison, it is necessary to ensure that the same amount of information is fed into the system for all the system sizes. This is achieved by keeping fixed the number M of projections and resolution of the images. As it will be shown in Sec. V B, the performance can be greatly enhanced when this information bottleneck is lifted and the amount of encoded information is varied. The first point to address is the trainability and generalization properties. In Fig. 5, we show the dependence of the training and testing errors on the generalization parameter λ. The curves in Fig. 5(a) are obtained assuming a full tomography and ideal measurements. Figure 5(b) instead presents the same results, but with imperfect measurements (see caption for more details). In Fig. 5(a),

the training error (dashed lines) drops to zero as $\lambda \rightarrow 0^+$; this is a manifestation of overfitting and indicates that, thanks to the high dimensionality of the quantum feature space, the system is able to completely fit the training data. Instead, the testing error (solid lines and markers) has a minimum value for some optimal value of λ, which depends on the dephasing rate γ (different markers denote different rates). For large enough values of λ the testing and training error curves eventually overlap. For increasing γ the minimum shifts to vanishing values of λ. A remarkable result is that the minimal testing error is very little affected by the dephasing rate. As shown in Fig. 5(b), the situation changes in the presence of imperfect measurements. Indeed, the minimum of the testing error is obtained for a finite value of λ even for large values of γ. Importantly, the minimum error increases with increasing dephasing noise. In Figs. 5(c) and 5(d), we report the dependence of the minimal testing error as a function of the number of spins N for increasing values of the dephasing rate. Again, Fig. 5(c) corresponds to ideal measurements, while curves in Fig. 5(d) are obtained under imperfect measurements. Figure 5(c) shows that the testing error diminishes as a
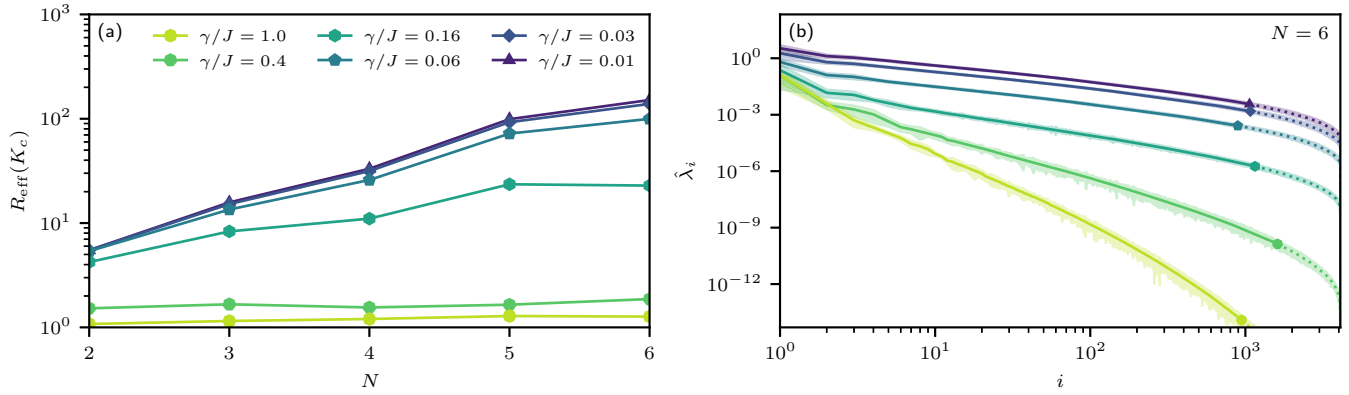
FIG. 6. (a) Kernel effective rank for the full tomography decoding as a function of the number of spins N and for different values of the dephasing rate γ. We have used the empirical representation of the kernel matrix on the training set. The results are averaged over the same numbers of disorder realizations as for Fig. 5. (b) Kernel empirical spectrum for the full tomography decoding, $N = 6$ spins, and different values of the dephasing rate γ. The markers correspond to the optimal generalization parameter λ. The curves have been obtained via the kernel empirical representation on the training set. Results are averaged over 25 disorder configurations. The filled area corresponds to twice the estimated standard error on the averaged value, using the same bootstrap method as for Fig. 5.

function of the number of spins and increases with dephasing rate. Note that also for very small dephasing rate the minimal testing error appears to saturate at large system sizes. This is hardly surprising as the input images have been preprocessed and considerably downsampled. This deliberate choice aims at making the task harder in order to gauge the expressivity of the machine without overloading the input information. As shown in Fig. 5(d), by considering imperfect measurements the role of dephasing is dramatically amplified.

As we have described in the analytical discussion in Sec. III, the quantum kernel spectrum allows us to assess the capacity of our model independently from the specific task one wants to achieve. Figure 6(a) shows the dependence of the quantum kernel's effective rank $R_{\text{eff}}(K_c)$ on the system size and noise strength. For vanishing values of the dephasing rate γ, we see that this figure of merit first increases exponentially with the number of spins before saturating. For increasing γ, the effective quantum kernel rank decreases, approaching 1 in the limit of very large γ. The same behavior is observed in the empirical spectrum in Fig. 6(b) as the noise rate is varied. For increasing values of γ, we observe a faster decrease of the empirical kernel eigenvalues as a function of the eigenvalue number. For comparison, we have indicated with markers the largest eigenvalue below the optimal generalization parameter. This gives a rough estimate of the number of kernel eigenfunctions required to correctly approximate the target function. Note that in the context of imperfect measurements the generalization parameter is bounded from below, and hence some of the kernel eigenfunctions becomes out of reach. This shows a clear link between the kernel eigenvalues and the expressivity of the machine. The results discussed relied on full tomography. As we have explained in Sec. IV, it is possible to design a simplified and less expensive measurement protocol based on a time multiplexing procedure where a set of local spin observables is measured at N_{rep} different times. The results obtained with such an approach are summarized in Fig. 7. As appears from Fig. 7(a), by increasing the number of repetitions N_{rep} the error diminishes. For small enough values of dephasing γ, the error converges to the

value in the ideal case of full tomography. For increasing γ, however, the saturating value departs from the ideal one given by full tomography, showing that the time multiplexing expressivity deteriorates more than that of the full tomography for larger noise. This trend is further elucidated in Fig. 7(b), where the difference between the time multiplexing error and the full-tomography error is reported as a function of the total number of measured observables. By increasing the number of spins and hence the dimension of the Hilbert space for a given dephasing rate, the required number of repetitions increases.

B. Optimizing the encoding

In this section we investigate an alternative encoding scheme for which the amount of information fed to the system scales with the system size. The embedding studied in the previous sections involved a set of $M = N_{\text{pulse}}$ random-projection features derived from the downsampled images. Here we derive a number $M = N \times N_{\text{pulse}}$ of such features and split them in N sequences of N_{pulse} features, which we use to drive the N sites. In particular, the driving sequences sent to different sites are unique, while for the previous encoding those sequences were proportional to each other. The new encoding procedure is presented in Fig. 8 in the form of its equivalent circuit. In Fig. 9 we report the evolution of the performances given by this new encoding as a function of the number of driving pulses. As the number of pulses rises the corresponding number of encoded features $M = N \times N_{\text{pulse}}$ increases and so does the amount of encoded information. The corresponding maximal testing accuracy reaches an optimum of 94.5% for $N_{\text{pulse}} = 3$. For large number of pulses N_{pulse} the performances drop. This effect is due to the fact that for such parameters the transformations yielded by the encoding are poorly adapted to the task at hand, i.e., the kernel and the target function become less aligned. Note that the performance is very sensitive to both the encoding method and the physical system parameters. While controlling the physical parameters might be hard, it appears that a careful design of the encoding procedure can significantly boost the
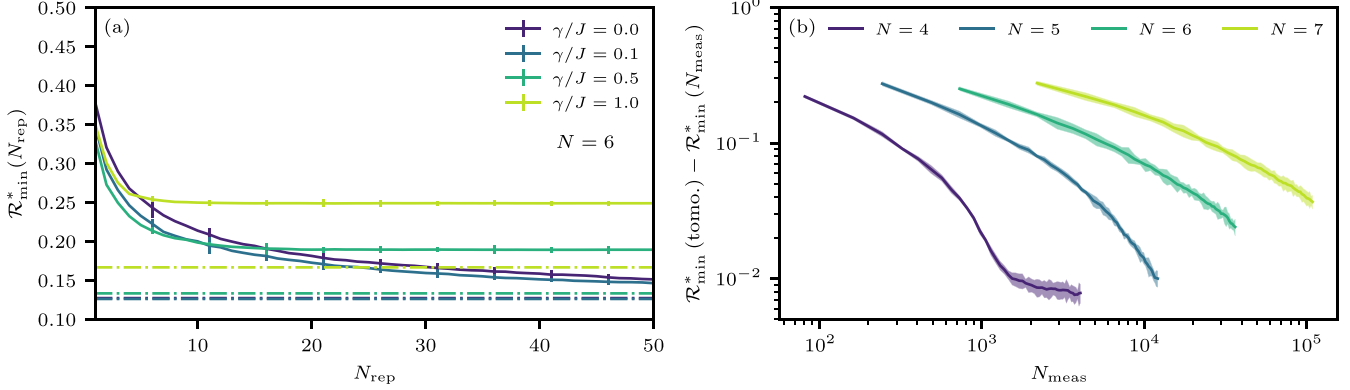
FIG. 7. (a) Minimal testing error as a function of the number of repetitions N_{rep} of all the local spin measurements for $N = 6$ spins and different values of the dephasing noise γ. The regularization parameter value is chosen to minimize the testing error. The horizontal dashed lines represent the errors obtained using a full tomography decoding for the corresponding dephasing rates. The results are averaged over 50 disorder realizations. (b) Difference between the testing error of the time multiplexing decoding and the one for full tomography as a function of the number $N_{\text{meas}} = 3NN_{\text{rep}}$ of local measurements performed for $N_{\text{rep}} = 50$, dephasing rate $\gamma/J = 0.01$, and different values of the number N of spins. For each realization of the disorder, the regularization parameter value is chosen to minimize the testing error. The results are averaged over 25 disorder realizations.

performances. This makes the research of tailored encoding procedures a promising avenue of research.

VI. CONCLUSION

In this paper, we have presented a quantum machine-learning model based on the quantum-kernel paradigm. Within the formalism of kernel theory, we have characterized the expressivity and generalization capacity of this model. We have linked the relevant figure of merits to the spectrum of the associated centered quantum kernel. In particular, we presented an upper bound on the generalization error involving the average purity of quantum states representing the data to classify. This upper bound shows that dissipation and decoherence act as a regularization for the quantum kernel machines. By considering an illustrating example of a driven-dissipative spin chain as the noisy quantum kernel machine, we have shown how the expressivity and generalization capacity are controlled by both the dephasing rate and experimental uncertainties on the measurements. Moreover, we have shown how the performances of the noisy quantum kernel machines are modified when the full-tomography measurement protocol

is replaced by a time multiplexing procedure requiring only local observables, and how the openness of the system mitigates the efficiency of this protocol. We observed a qualitative improvement in the processing performance of our model when going from a scenario where the system is fed a constant amount of information to one where the inputs are encoded at a finite information rate that scales extensively with the system size. How to design tailored encoding strategies able to harness the full power of quantum kernel machines remains an open question. In particular, investigating encoding schemes that would allow one to inject information at a rate scaling exponentially in the system size seems promising. The concepts presented here and the unavoidable role of the decoherence in any realistic physical system are relevant for a wide range of quantum machine-learning models, ranging from quantum extreme-learning machines to quantum neural networks.
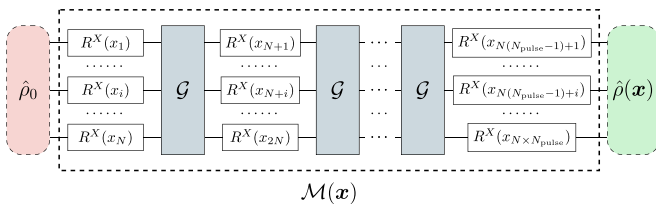
FIG. 8. Equivalent circuit of the encoding with the information bottleneck removed. Instead of injecting a single random projection feature x_i per time step as represented in Fig. 3, where a total number of $M = N_{\text{pulse}}$ random projection features are fed into the kernel machine, here we inject N random projection features in each time step, with a total number of $M = N \times N_{\text{pulse}}$ random projection features injected by the end of the encoding sequence.

APPENDIX A: EXPRESSIVITY AND GENERALIZATION FOR NOISY QUANTUM KERNELS

1. Expressivity and kernel effective rank

To measure the ability of a kernel k to learn a function $y(\boldsymbol{x})$, we have introduced in Eq. (27) the *kernel target alignment* $A(k, y)$. We then defined the *kernel effective rank* R_{eff} by considering a set of orthonormal basis functions $\{g_i\}$, that
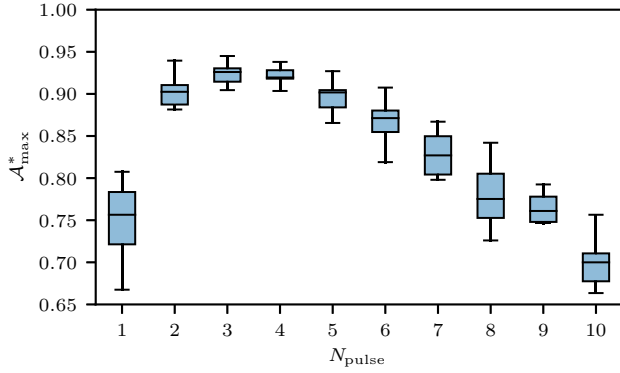
FIG. 9. Optimal testing accuracy as a function of the number of driving pulses, for the encoding presented in Fig. 8, $N = 6$ spins, and $\gamma/J = 0.01$. The regularization parameter is chosen to maximize the testing accuracy. Note that for this encoding the number of features M yielded by the preprocessing is $M = N \times N_{\text{pulse}}$. The results are shown for ten realizations of the disorder on both the preprocessing and the system parameters. The maximal accuracy obtained is 94.5% for $N_{\text{pulse}} = 3$. The boxes extend from the first to the third quartile of the distributions, the middle line indicates the median, and the whiskers indicate the extreme values of the distributions.

gives the following equalities:

$$
\begin{aligned}
\sqrt{R_{\text{eff}}(k)} &= \sum_j A(k, g_j) \\
&= \frac{1}{\left(\sum_i \lambda_i^2\right)^{1/2}} \sum_j \sum_i \lambda_i \mathbb{E}_p[\psi_i(\boldsymbol{x}) g_j(\boldsymbol{x})]^2 \\
&= \frac{1}{\left(\sum_i \lambda_i^2\right)^{1/2}} \sum_i \lambda_i \mathbb{E}_p[\psi_i(\boldsymbol{x})^2] \\
&= \frac{\sum_i \lambda_i}{\left(\sum_i \lambda_i^2\right)^{1/2}}.
\end{aligned}
\tag{A1}
$$

Note that the final expression concerns only the spectrum of the kernel and is independent of the choice of the basis functions $\{g_i\}$. From the Cauchy-Schwarz inequality, we have

$$
R_{\text{eff}}(k) \leqslant |\{\lambda_i \neq 0\}|,
\tag{A2}
$$

where the equality is attained if and only if all nonzero eigenvalues of the kernel are equal. Therefore, it provides information about the flatness of the spectrum of the kernel. Given a training sample of size N_{train}, the kernel spectrum can be empirically computed using the $N_{\text{train}} \times N_{\text{train}}$ kernel matrix $\boldsymbol{K}$ associated to the kernel k, the entries of which are $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The eigenvalues λ_i of the kernel k can then be approximated by those of the matrix $\boldsymbol{K}/N_{\text{train}}$ [57]. For the centered quantum kernel k_c with the associated kernel matrix $\boldsymbol{K}_c$, we can compute the effective rank empirically as

$$
\sqrt{R_{\text{eff}}(\boldsymbol{K}_c)} = \frac{\text{Tr}[\boldsymbol{K}_c]}{\sqrt{\text{Tr}[\boldsymbol{K}_c^2]}} = \frac{\sum_i \hat{\lambda}_i}{\sqrt{\sum_i \hat{\lambda}_i^2}},
\tag{A3}
$$

where the $\hat{\lambda}_i$ are the empirical eigenvalues [58].

The numerator can be expressed using the empirical kernel eigenobservables. In order to keep light notations we use the same notations as in the main text, i.e., the empirical kernel eigenobservables are denoted $\hat{E}_i$. Whether this notation refers to the exact or the empirical observable should be clear from the context. We have for the numerator

$$
\sum_i \hat{\lambda}_i = \mathbb{E}_{\hat{p}}\left[\sum_i \delta\langle\hat{E}_i\rangle_{\boldsymbol{x}}^2\right] = \mathbb{E}_{\hat{p}}\left[\sum_i \text{Tr}[\delta\hat{\rho}(\boldsymbol{x})\hat{E}_i]^2\right].
\tag{A4}
$$

Since $\text{Tr}[\delta\hat{\rho}(\boldsymbol{x})] = 0$, $\delta\hat{\rho}(\boldsymbol{x})$ can be decomposed onto the eigenobservable basis $\{\hat{E}_i\}$ through the expression

$$
\delta\hat{\rho}(\boldsymbol{x}) = \sum_i \text{Tr}[\delta\hat{\rho}(\boldsymbol{x})\hat{E}_i]\hat{E}_i.
\tag{A5}
$$

Consequently, the squared Hilbert-Schmidt norm reads

$$
\text{Tr}[\delta\hat{\rho}(\boldsymbol{x})^2] = \sum_i \text{Tr}[\delta\hat{\rho}(\boldsymbol{x})\hat{E}_i]^2.
\tag{A6}
$$

Equation (A4) therefore becomes

$$
\begin{aligned}
\sum_i \hat{\lambda}_i &= \mathbb{E}_{\hat{p}}[\text{Tr}[\delta\hat{\rho}(\boldsymbol{x})^2]] = \text{Tr}[\mathbb{E}_{\hat{p}}[\{\hat{\rho}(\boldsymbol{x}) - \mathbb{E}_{\hat{p}}[\hat{\rho}(\boldsymbol{x})]\}^2]] \\
&= \mathbb{E}_{\hat{p}}[\text{Tr}[\hat{\rho}(\boldsymbol{x})^2]] - \text{Tr}[\mathbb{E}_{\hat{p}}[\hat{\rho}(\boldsymbol{x})]^2],
\end{aligned}
\tag{A7}
$$

giving Eq. (32) in the main text (as the same relation holds between the true eigenvalues λ_i and the distribution p). This quantity can also be written in terms of the measured observables (note that $\mathcal{O} = \mathcal{B}$ for a quantum kernel)

$$
\begin{aligned}
\sum_i \hat{\lambda}_i &= \frac{\text{Tr}[\boldsymbol{K}_c]}{N_{\text{train}}} = \frac{1}{N_{\text{train}}} \sum_i k_c(\boldsymbol{x}_i, \boldsymbol{x}_i) \\
&= \frac{1}{N_{\text{train}}} \sum_i \sum_k \delta\phi_k(\boldsymbol{x}_i)\delta\phi_k(\boldsymbol{x}_i) \\
&= \sum_k \left(\frac{1}{N_{\text{train}}} \sum_i \delta\phi_k(\boldsymbol{x}_i)\delta\phi_k(\boldsymbol{x}_i)\right) \\
&= \sum_k \mathbb{E}_{\hat{p}}[\delta\phi_k(\boldsymbol{x})^2] = \sum_k \mathbb{E}_{\hat{p}}[\delta\langle\hat{O}_k\rangle_{\boldsymbol{x}}^2] \\
&= \sum_k \text{Var}_{\hat{p}}[\langle\hat{O}_k\rangle_{\boldsymbol{x}}].
\end{aligned}
\tag{A8}
$$

Similarly, in the denominator of Eq. (A3), we get

$$
\begin{aligned}
\sum_i \hat{\lambda}_i^2 &= \frac{\text{Tr}[\boldsymbol{K}_c^2]}{N_{\text{train}}^2} \\
&= \frac{1}{N_{\text{train}}^2} \sum_{i,j} k_c(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \\
&= \frac{1}{N_{\text{train}}^2} \sum_{i,j} \left(\sum_k \delta\phi_k(\boldsymbol{x}_i)\delta\phi_k(\boldsymbol{x}_j)\right)^2 \\
&= \sum_{k,l} \left(\frac{1}{N_{\text{train}}} \sum_i \delta\phi_k(\boldsymbol{x}_i)\delta\phi_l(\boldsymbol{x}_i)\right)^2 \\
&= \sum_{k,l} \mathbb{E}_{\hat{p}}[\delta\langle\hat{O}_k\rangle_{\boldsymbol{x}}\delta\langle\hat{O}_l\rangle_{\boldsymbol{x}}]^2 \\
&= \sum_{k,l} \text{Cov}_{\hat{p}}[\langle\hat{O}_k\rangle_{\boldsymbol{x}}, \langle\hat{O}_l\rangle_{\boldsymbol{x}}]^2.
\end{aligned}
\tag{A9}
$$

Finally, we get the general expression

$$\sqrt{R_{\text{eff}}(\boldsymbol{K}_c)} = \frac{\sum_{i=1}^{P} \text{Var}_{\hat{p}}[\langle \hat{O}_i \rangle_{\boldsymbol{x}}]}{\left(\sum_{i,j=1}^{P} \text{Cov}_{\hat{p}}[\langle \hat{O}_i \rangle_{\boldsymbol{x}}, \langle \hat{O}_j \rangle_{\boldsymbol{x}}]^2 \right)^{\frac{1}{2}}}. \qquad (A10)$$

Note that this relation also holds for the true (nonempirical) effective rank $R_{\text{eff}}(k_c)$ provided that the variances and the covariances are taken with respect to the true probability distribution p instead of the empirical one $\hat{p}$.

2. Generalization and Rademacher complexity

Here we give the detailed derivation of Eq. (38) using methods of statistical learning theory applied to the specific case of a noisy centered quantum kernel [51]. In the standard setup of statistical learning, the inputs $\boldsymbol{x} \in \mathcal{X}$ are considered as a random variable following a probability distribution $p(\boldsymbol{x})$. We define the target function $y : \mathcal{X} \mapsto \mathcal{Y}$ that assigns to each input its right label. We will consider the case of a binary classification, for which $\mathcal{Y} = \{-1, 1\}$. In practice the true distribution p of the inputs is unknown and during the training we only have access to a finite training dataset $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \mid i = 1, \dots, N_{\text{train}}\}$. The elements of the dataset are considered as realizations of a set of independent and identically distributed random variables following p. The empirical distribution associated to this training set is given by

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \delta(\boldsymbol{x} - \boldsymbol{x}_i). \qquad (A11)$$

We rely on this empirical distribution to evaluate expectations of any function $f(\boldsymbol{x})$, namely,

$$\mathbb{E}_p[f(\boldsymbol{x})] = \int_{\mathcal{X}} f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}. \qquad (A12)$$

The expectation value is approximated by its empirical counterpart:

$$\mathbb{E}_{\hat{p}}[f(\boldsymbol{x})] = \int_{\mathcal{X}} f(\boldsymbol{x}) \hat{p}(\boldsymbol{x}) d\boldsymbol{x} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} f(\boldsymbol{x}_i). \qquad (A13)$$

A common question in statistical learning is to know how a model trained on a given set of data will perform on any other set of unseen data. For a binary classification task with balanced data one can use the accuracy as a measure of the model performance. Given a trial function $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$ that has been optimized using the training set $\mathcal{S}$, we define the corresponding prediction function as $\tilde{f}(\boldsymbol{x}) = \text{sign}[f(\boldsymbol{x})]$. An input $\boldsymbol{x}$ is correctly classified if $\tilde{f}(\boldsymbol{x}) = y(\boldsymbol{x})$. The true accuracy $\mathcal{A}^*(f)$ is defined as the probability that any input in $\mathcal{X}$ is correctly classified by $\tilde{f}$:

$$\mathcal{A}^*(f) = \mathbb{E}_p\big[\mathbb{1}_{y(\boldsymbol{x}) = \tilde{f}(\boldsymbol{x})}\big] = \mathbb{E}_p\big[\mathbb{1}_{y(\boldsymbol{x}) f(\boldsymbol{x}) \geqslant 0}\big] \qquad (A14)$$

$$= 1 - \mathbb{E}_p\big[\mathbb{1}_{y(\boldsymbol{x}) f(\boldsymbol{x}) \leqslant 0}\big] = 1 - \mathcal{R}^*(f), \qquad (A15)$$

where we define the risk (also called error or inaccuracy) as $\mathcal{R}^*(f) = 1 - \mathcal{A}^*(f)$. The corresponding empirical quantities $\mathcal{A}(f)$ and $\mathcal{R}(f)$ are defined in an analogous way using the empirical distribution $\hat{p}$ instead of p. The ability to perform well on new data is measured by the generalization error:

$$\mathcal{E}(f) = \mathcal{R}^*(f) - \mathcal{R}(f). \qquad (A16)$$

Statistical learning theory provides probabilistic upper bounds on the generalization error depending on the type of task at hand and on the specific model used to tackle it. In order to find such an upper bound for a binary classification task, it is convenient to consider a relaxed version of the risk, the η-margin risk $\mathcal{R}_\eta(f)$ defined in the main text. The upper bound on the generalization properties involves the empirical Rademacher complexity of a class of trial functions $\mathcal{H}$ with respect to the training sample $\mathcal{S}$. It is defined as

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{H}} \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \sigma_i f(\boldsymbol{x}_i) \right] \qquad (A17)$$

where $\boldsymbol{\sigma}$ is a vector of Rademacher variables that are discrete, independent, and identically distributed following a uniform law over $\{-1, 1\}$. The Rademacher complexity measures the ability of a hypothesis class $\mathcal{H}$ to fit noise, and as such it is a measure of the expressivity of $\mathcal{H}$. We now give an upper bound on the generalization error (Theorem 5.8 in [51]).

Theorem 1. Let $\mathcal{H}$ be a set of trial functions and $\eta > 0$. Then $\forall \delta > 0$, with probability at least $1 - \delta$, and we have $\forall f \in \mathcal{H}$:

$$\mathcal{R}^*(f) \leqslant \mathcal{R}_\eta(f) + \frac{2}{\eta} \mathfrak{R}_{\mathcal{S}}(\mathcal{H}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2 N_{\text{train}}}}.$$

This upper bound can be specialized to the case of kernel methods where the hypothesis class is the RKHS of a kernel k. In this the Rademacher complexity is upper bounded by a quantity that depends only on the trace of the empirical kernel matrix $\boldsymbol{K}$ (Theorem 6.12 in [51]).

Theorem 2. Let $\mathcal{H}$ by the RKHS associated to a given kernel k. For $\Lambda \geqslant 0$ consider the set of hypothesis functions $\mathcal{H}_\Lambda = \{f : x \mapsto \boldsymbol{w}^T \delta \boldsymbol{\phi}(x), \quad \|\boldsymbol{w}\|^2 \Lambda \leqslant 1\} \subseteq \mathcal{H}$. Then we have

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{H}_\Lambda) \leqslant \frac{1}{N_{\text{train}}} \sqrt{\frac{\text{Tr}[\boldsymbol{K}]}{\Lambda}}.$$

Injecting this result in the previous upper bound, we get the desired result. In particular, using the centered noisy quantum kernel k_c and Eq. (A7), we get Eq. (38).

APPENDIX B: INTERCEPT AND KERNEL CENTERING

The initial optimization problem is to find a weight vector $\boldsymbol{w} = (b, w_1, \dots, w_P)^T$ that minimizes the regularized loss function of Eq. (8). We slightly change our notation and drop the first constant term of the embedding map $\boldsymbol{\phi}(\boldsymbol{x})$ to explicitly separate the bias b from the weight $\boldsymbol{w}$, so that the loss can be rewritten:

$$\mathcal{L}(\boldsymbol{w}, b \mid \mathcal{S}) = \frac{1}{2 N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} (y_i - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i) - b)^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2. \qquad (B1)$$

The optimal intercept b is found by imposing $\frac{\partial \mathcal{L}}{\partial b} = 0$. The solution reads

$$b^* = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} y_i - \boldsymbol{w}^T \left(\frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \boldsymbol{\phi}(x_i) \right). \quad \text{(B2)}$$

We see that the optimal intercept consists of two terms: one that has the effect of centering the labels, while the other centers the features. Assuming the dataset we use is balanced, we have $\sum_{i=1}^{N_{\text{train}}} y_i \simeq 0$. Plugging the optimal intercept back into the previous regularized loss function, we get a new effective loss:

$$\mathcal{L}^*(\boldsymbol{w} \mid \mathcal{S}) = \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} (y_i - \boldsymbol{w}^T \{\boldsymbol{\phi}(x_i) - \mathbb{E}_{\hat{p}}[\boldsymbol{\phi}]\})^2$$
$$+ \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2. \quad \text{(B3)}$$

If the data are not balanced one can simply replace the labels y_i by their centered counterpart $y_i' = y_i - \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} y_i$ such that $\sum_{i=1}^{N_{\text{train}}} y_i' = 0$. Note that this might lead to issues when using the accuracy metric with unbalanced labels. This issue can be fixed, e.g., by changing the metric used for a balanced one. Thus, working with the quantum kernel without regularizing the intercept term is equivalent to working with the centered kernel and centered labels.

APPENDIX C: TIME MULTIPLEXING AND MODEL EXPRESSIVITY

The maximal class of trial functions $\mathcal{H}_{\text{full}}$ (see Sec. III) associated to a given embedding is obtained by performing a complete tomography of the embedded quantum states $\hat{\rho}(\boldsymbol{x})$ right after the end of the encoding procedure. The system evolution after time τ according to a Lindblad master equation with a constant Hamiltonian and dissipator for a given duration δt_m can be expressed into a set of Kraus operators $\{\hat{W}_i\}$ [34] satisfying

$$\sum_i \hat{W}_i^\dagger \hat{W}_i = \hat{\mathbb{1}}. \quad \text{(C1)}$$

The evolved density matrices $\hat{\rho}(\boldsymbol{x}; \delta t_m)$ are given by

$$\hat{\rho}(\boldsymbol{x}; \delta t_m) = \sum_i \hat{W}_i \hat{\rho}(\boldsymbol{x}) \hat{W}_i^\dagger. \quad \text{(C2)}$$

In the Heisenberg picture, the observables evolve in time following an adjoint master equation [34]. Hence, we can see the nonunitary evolution of the open quantum system as a simple change in the set of observables that are measured on the state $\hat{\rho}(\boldsymbol{x})$. Suppose that we want to measure observables from the orthonormal basis introduced in Sec. II B after the previous evolution. We define the $(P+1) \times (P+1)$ matrix $\boldsymbol{\Xi}$ the elements of which are

$$\Xi_{kl} = \text{Tr}\left[\sum_i \hat{W}_i^\dagger \hat{O}_k \hat{W}_i \hat{O}_l \right]. \quad \text{(C3)}$$

The measurement at time $\tau + \delta t_m$ of the observable $\hat{O}_l$ can now be expressed using the decomposition in Eq. (44) and the elements of $\boldsymbol{\Xi}$ as

$$\text{Tr}[\hat{\rho}(\boldsymbol{x}; \delta t_m) \hat{O}_l] = \frac{1}{2^N} \left\{ \Xi_{0l} + \sum_k \text{Tr}[\hat{\rho}(\boldsymbol{x}) \hat{O}_k] \Xi_{kl} \right\}. \quad \text{(C4)}$$

Thus the embedding map $\boldsymbol{\phi}(\boldsymbol{x})$ is transformed by the nonunitary evolution during δt_m and becomes

$$\boldsymbol{\phi}(\boldsymbol{x}; \delta t_m) = \boldsymbol{\Xi} \boldsymbol{\phi}(\boldsymbol{x}). \quad \text{(C5)}$$

Assuming we only make measurements on a subset of the basis $\{\hat{O}_i\}$, then we can write for the feature vector

$$\boldsymbol{\phi}(\boldsymbol{x}; \delta t_m) = \boldsymbol{D} \boldsymbol{\Xi} \boldsymbol{\phi}(\boldsymbol{x}), \quad \text{(C6)}$$

where $\boldsymbol{D}$ is a diagonal $(P+1) \times (P+1)$ matrix the diagonal entries i of which are 1 if $\hat{O}_i$ is measured and zero otherwise. When we repeat the measurements at different times, we can stack the previous vectors at each time step. For N_{rep} repetitions, we denote $\boldsymbol{\Lambda}$ the $N_{\text{rep}}(P+1) \times (P+1)$ matrix of the form

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{D}\boldsymbol{\Xi} \\ \boldsymbol{D}\boldsymbol{\Xi}^2 \\ \vdots \\ \boldsymbol{D}\boldsymbol{\Xi}^{N_{\text{rep}}} \end{pmatrix}. \quad \text{(C7)}$$

The final vector reads

$$\tilde{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{\Lambda} \boldsymbol{\phi}(\boldsymbol{x}). \quad \text{(C8)}$$

Hence, performing repeated measurements in between nonunitary evolutions amounts to performing a restricted number of measurements on the encoded states $\hat{\rho}(\boldsymbol{x})$ at time τ. This implies that the time multiplexing decoding lowers the model expressivity. The difference between the models obtained from the full tomography and the time multiplexing decoding is encapsulated in the matrix $\boldsymbol{\Lambda}$.

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, edited by F. Bach, Adaptive Computation and Machine Learning Series (MIT, Cambridge, MA, 2016).

[3] E. Strubell, A. Ganesh, and A. McCallum, Energy and policy considerations for deep learning in NLP, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, 2019), pp. 3645–3650.

[4] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, Recent advances in physical reservoir computing: A review, Neural Networks **115**, 100 (2019).

[5] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, Extreme learning machine: Theory and applications, Neurocomputing Neural Networks **70**, 489 (2006).

[6] A. Opala, S. Ghosh, T. C.H. Liew, and M. Matuszewski, Neuromorphic Computing in Ginzburg-Landau Polariton-Lattice Systems, Phys. Rev. Appl. **11**, 064029 (2019).

[7] Z. Denis, I. Favero, and C. Ciuti, Photonic Kernel Machine Learning for Ultrafast Spectral Analysis, Phys. Rev. Appl. **17**, 034077 (2022).

[8] D. Ballarini, A. Gianfrate, R. Panico, A. Opala, S. Ghosh, L. Dominici, V. Ardizzone, M. De Giorgi, G. Lerario, G. Gigli, T. C. H. Liew, M. Matuszewski, and D. Sanvitto, Polaritonic neuromorphic computing outperforms linear classifiers, Nano Lett. **20**, 3506 (2020).

[9] D. Pierangeli, G. Marcucci, and C. Conti, Photonic extreme learning machine by free-space optical propagation, Photon. Res. **9**, 1446 (2021).

[10] T. Hofmann, B. Schölkopf, and A. J. Smola, Kernel methods in machine learning, Ann. Stat. **36**, 1171 (2008).

[11] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems* (Curran, New York, 2018), Vol. 31.

[12] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).

[13] V. Dunjko and H. J. Briegel, Machine learning and artificial intelligence in the quantum domain: A review of recent progress, Rep. Prog. Phys. **81**, 074001 (2018).

[14] M. Schuld, I. Sinayskiy, and F. Petruccione, The quest for a quantum neural network, Quant. Info. Proc. **13**, 2567 (2014).

[15] D. Marković and J. Grollier, Quantum neuromorphic computing, Appl. Phys. Lett. **117**, 150501 (2020).

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

[17] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors, arXiv:1802.06002.

[18] D. K. Park, C. Blank, and F. Petruccione, The theory of the quantum kernel-based binary classifier, Phys. Lett. A **384**, 126422 (2020).

[19] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[20] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, Quantum embeddings for machine learning, arXiv:2001.03622.

[21] T. Hubregtsen, D. Wierichs, E. Gil-Fuster, P.-J. H. S. Derks, P. K. Faehrmann, and J. J. Meyer, Training quantum embedding kernels on near-term quantum computers, Phys. Rev. A **106**, 042431 (2022).

[22] M. Schuld, Supervised quantum machine learning models are kernel methods, arXiv:2101.11020.

[23] T. Kusumoto, K. Mitarai, K. Fujii, M. Kitagawa, and M. Negoro, Experimental quantum kernel trick with nuclear spins in a solid, npj Quantum Inf. **7**, 94 (2021).

[24] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. **17**, 1013 (2021).

[25] Y. Wu, B. Wu, J. Wang, and X. Yuan, Provable advantage in quantum phase learning via quantum kernel alphatron, arXiv:2111.07553.

[26] J. M. Kübler, S. Buchholz, and B. Schölkopf, The inductive bias of quantum kernels, arXiv:2106.03747.

[27] R. Shaydulin and S. M. Wild, Importance of kernel bandwidth in quantum machine learning, Phys. Rev. A **106**, 042407 (2022).

[28] X. Wang, Y. Du, Y. Luo, and D. Tao, Towards understanding the power of quantum kernels in the NISQ era, Quantum **5**, 531 (2021).

[29] K. Bartkiewicz, C. Gneiting, A. Černoch, K. Jiráková, K. Lemr, and F. Nori, Experimental kernel-based quantum machine learning in finite feature space, Sci. Rep. **10**, 12356 (2020).

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

[31] K. Fujii and K. Nakajima, Harnessing Disordered-Ensemble Quantum Dynamics for Machine Learning, Phys. Rev. Appl. **8**, 024030 (2017).

[32] H. Xu, T. Krisnanda, W. Verstraelen, T. C. H. Liew, and S. Ghosh, Superpolynomial quantum enhancement in polaritonic neuromorphic computing, Phys. Rev. B **103**, 195302 (2021).

[33] In the case of multivariate functions the variables against which the expectation is taken are indicated in the subscript.

[34] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University, New York, 2007).

[35] J. A. K. Suykens and J. Vandewalle, Least squares support vector machine classifiers, Neural Proc. Lett. **9**, 293 (1999).

[36] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2013).

[37] Note that in certain scenarios one may prefer to exclude the first component of $\bar{w}$, which is a constant intercept term, in the regularization. Then it suffices to replace the $(1,1)$ entry of $\mathbb{1}$ by zero in Eq. (9) to obtain the optimal weights. This is equivalent to using a centered kernel, as discussed in Appendix B.

[38] S. Jerbi, L. J. Fiderer, H. P. Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, Quantum machine learning beyond kernel methods, arXiv:2110.13162.

[39] The probability measure p on the input space $\mathcal{X}$ is important here as it determines the scalar product on the space of real-valued functions on $\mathcal{X}$ through $\langle f, g \rangle = \mathbb{E}_{\boldsymbol{x} \sim p}[f(\boldsymbol{x})g(\boldsymbol{x})]$. The reproducing property, crucial to link the RKHS and its kernel, relies on such a well-defined scalar product.

[40] V. I. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge Studies in Advanced Mathematics (Cambridge University, Cambridge, England, 2016).

[41] 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).

[42] For a closed system, this quantum kernel can be directly evaluated through measurement [18] and the trial function can be expressed in terms of the quantum kernel and optimized in an equivalent way.

[43] B. Bordelon, A. Canatar, and C. Pehlevan, Spectrum dependent learning curves in kernel regression and wide neural networks, arXiv:2002.02561.

[44] A. Canatar, B. Bordelon, and C. Pehlevan, Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks, Nat. Commun. **12**, 2914 (2021).

[45] K. Nakaji, H. Tezuka, and N. Yamamoto, Quantum-enhanced neural networks in the neural tangent kernel framework, arXiv:2109.03786.

[46] J. Liu, F. Tacchino, J. R. Glick, L. Jiang, and A. Mezzacapo, Representation learning via quantum neural tangent kernels, arXiv:2111.04225.

[47] 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).

[48] T. Hastie and J. Zhu, Comment: Support vector machines with applications, Stat. Sci. **21**, 352 (2006).

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

[50] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, On kernel target alignment, in *Innovations in Machine Learning: Theory and Applications*, Studies in Fuzziness and Soft Computing, edited by D. E. Holmes and L. C. Jain (Springer-Verlag, Berlin, 2006), pp. 205–256.

[51] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning* (MIT, Cambridge, MA, 2018).

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

[53] In total, the results presented in this paper required approximately 800 000 scalar hours ($\approx$90 years) of computation and 5 TB of storage on the acknowledged French National High Performance Computing facility (GENCI). During the simulations, we used approximately up to 30 000 cores simultaneously.

[54] This can be explicitly expressed as a D-deep circuit via Trotterization as $\mathcal{G} = [\prod_{i=1}^{N} R_i^Z(-\frac{2\Delta_i\Delta t}{D}) \prod_{\langle i,j \rangle} \prod_{K \in \{X,Y,Z\}} R_{ij}^{KK}(\frac{2J_{ij}^K \Delta t}{D})]^D + O(J_0^2 \Delta t^2 / D^2)$.

[55] A. Czerwinski, Quantum state tomography with informationally complete POVMs generated in the time domain, Quant. Info. Proc. **20**, 105 (2021).

[56] A. Di Lorenzo, Sequential Measurement of Conjugate Variables as an Alternative Quantum State Tomography, Phys. Rev. Lett. **110**, 010404 (2013).

[57] C. Williams and M. Seeger, Using the Nyström method to speed up kernel machines, in *Advances in Neural Information Processing Systems 13* (MIT, Cambridge, MA, 2001), pp. 682–688.

[58] The hat symbol is used for estimators such as the empirical eigenvalues, as is customary in statistical theory. The hat must not be confused with the one used for the quantum operators.