# Photonic Kernel Machine Learning for Ultrafast Spectral Analysis

Zakari Denis, Ivan Favero, and Cristiano Ciuti*
*Université Paris Cité, CNRS, Laboratoire Matériaux et Phénomènes Quantiques, 75013 Paris, France*

(Received 8 November 2021; accepted 17 February 2022; published 31 March 2022)

We introduce photonic kernel machines, a scheme for ultrafast spectral analysis of noisy radio-frequency signals from single-shot optical intensity measurements. The approach combines the versatility of machine learning and the speed of photonic hardware to reach unprecedented throughput rates. We theoretically describe some of the key underlying principles, and then numerically illustrate the performance achieved in a photonic lattice-based implementation. We apply the technique both to picosecond pulsed radio-frequency signals, on energy-spectral-density estimation and a shape-classification task, and to continuous signals, on a frequency-tracking task. The optical computing scheme presented is resilient to noise while requiring minimal control on the photonic-lattice parameters, making it readily implementable in realistic state-of-the-art photonic platforms.

DOI: 10.1103/PhysRevApplied.17.034077

## I. INTRODUCTION

As a result of its intrinsically faster time-scales, photonics was very quickly seen as a promising tool to outperform integrated electronics in terms of data processing rates [1]. In this perspective, optical setups ranging from matrix-vector multipliers [2], function convolvers [3], and discrete Fourier-transforming processors [4] to non-von-Neumann parallel digital processors [5,6] were proposed. Although these ideas quickly became outdated as a consequence of the fast rise of silicon-based electronic processors, the latter have started to exhibit some of their limitations. In particular, the emergence of machine-learning applications operating on ever increasing amounts of data, involving deeper and deeper neural-network architectures with an increasing degree of complexity [7], has led to a situation where the progress of available digital processor technology no longer keeps pace with the demand in computing capabilities [8]. Moreover, this trend has gone hand in hand with an increase in the consumption of computational and energy resources, casting doubt on its sustainability [9].

This context has stimulated very interesting proposals aiming at surrogating the realization of specific tasks that are computationally and energetically very demanding to very specialized (electro-)optical devices. This process, reminiscent of current trends in hardware-acceleration technologies, such as graphical (GPUs) and tensor processing units (TPUs), has already led to commercially available optical coprocessors [10–13]. In recent years, this approach has been scaled to deep architectures [14] and has proven spectacularly powerful in the field of computer vision, by exploiting diffraction [15] or by optical implementations of convolutional neural networks, both in free-space [16] and on chip [17]. Such architectures are able to extract increasingly abstract representations [18] of the input images fed into the network by subsequent applications of pooled optically operated linear convolutions.

Standard neural-network-based machine-learning schemes present roughly the following architecture. The digital data to be processed are transformed by a series of consecutive layers that consist of a parametrized affine transformation followed by the pointwise application of an elementary nonlinear activation function. Such a composition of parametrized functions is then expected to approximate some target function of the input upon a proper training process involving the optimization of the parameters in order to minimize the error of the model over some set of training examples. While this sequential architecture is ideally suited for standard processors, which offer arbitrary levels of programmability by design, the number of parameters to be addressed during the training process is in practice a hurdle to flexible optical implementations, having progressed from roughly 63,000 in the paradigmatic LeNet-5 convolutional neural network [19] to several tens of millions in its present-day counterparts [7].

Therefore, machine-learning paradigms that relax the above deep-architecture constraints, such as extreme-learning machines [20,21], echo-state networks, or reservoir computing [22–31], have inspired theoretical proposals and experimental realizations in a variety of physical settings, in free-beam optics [32–34], integrated photonics [35,36], memristors [37,38], and beyond [39–41]. In the context of optics, many fruitful configurations have been investigated, such as delay-line-based setups

*cristiano.ciuti@u-paris.fr

[42–53], nonlinear polariton lattices [54–56], and systems combining linear light scattering and the nonlinearity of the measurement [34,57,58]. Beyond liquid-state machines and echo-state networks, several proposals have been put forward to reconcile general-purpose machine learning and distributed memory architectures. This is the case of neuromorphic approaches, in particular in the field of spintronics [59–63] and memristors [64–77], notably works articulated around spike-based machine-learning schemes [78–83]. Such architectures come with their own technical difficulties. Because they rely on co-located memory and processing resources, standard optimization algorithms can prove impractical, although optimization procedures have recently been put forward to overcome this obstacle by meeting their peculiar hardware design [84–88].

Physical machine-learning paradigms open prospects to face the two shortcomings of software machine learning: the limited processing throughput, and the handling of analog data that cannot be suitably interfaced with digital processors. Instances of the latter problem are, for example, situations where the input data to be analyzed are supplied at a rate too high to be properly sampled in real time. This also occurs when the data are intrinsically analog, or when direct measurement processes add noise or perturb the system being measured. The best example of this is provided by genuinely quantum tasks with no classical counterparts, involving quantum inputs [89]. This has stimulated many original works, ranging from quantum metrology [90,91] and quantum state control [92–94] to classical image-recognition tasks with quantum hardware [95], under the name of quantum neuromorphic computing [96].

Here we take a different path and propose to implement a distinct machine-learning paradigm, namely kernel machines [97,98], in the world of photonics. The approach appears to circumvent the two shortcomings listed above. We describe photonic kernel machines as well as the associated theoretical framework under very general assumptions. We explore the links between key concepts in support-vector-machine (SVM) theory and those of the machine proposed here. We show that, in contrast to general reservoir-computing schemes, knowledge about the underlying internal representations may be revealed from measurable data, providing direct understanding of the learning process of actual hardware. By introducing a realistic physical model for a photonic kernel machine based on a two-dimensional lattice of coupled linear optical cavities, we numerically examine the performance of photonic kernel machines on regression and classification tasks involving ultrafast spectral analysis of analog noisy radio-frequency (rf) signals, which are imprinted on an optical carrier wave.

This paper is organized as follows. After presenting some general concepts of kernel-machine theory in Sec. II, photonic kernel machines are introduced theoretically in Sec. III; their learning mechanism is discussed therein. A model for a physical implementation based on a photonic lattice is then described in Sec. IV. This model is simulated numerically in Sec. V and applied to the ultrafast spectral analysis of noisy rf signals from single-shot optical intensity measurements. Finally, conclusions are drawn in Sec. VI.

## II. GENERAL FRAMEWORK

Supervised machine-learning problems can be formulated as an optimization problem in which one tries to best approximate a *target* quantity $\mathbf{y} = f(\mathbf{x})$ for some given input vector $\mathbf{x} = [x_1, x_2, \ldots]^T$ with a parametrized function $\hat{f}$. The input data are distributed according to some unknown distribution $p(\mathbf{x})$ from which only a restricted set of examples $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$ is known.

More precisely, in *regression* problems one seeks the optimal parameters that make $\hat{f}$ the best fit for the known examples by approximating their unknown true functional dependence $f$. At variance, in classification problems one aims to determine a predictor $\hat{f}$ that best associates to any given input $\mathbf{x}$ a set of labels $\mathbf{y}$ that characterizes its belonging to one or more classes. These approximation problems are expressed in practice as the minimization of some cost function $J$ with respect to the parameters of the model to be trained, for a given set of examples.

A specific choice of parametrization defines the architecture of the model. Many such architectures exist, ranging from shallow models [99], such as SVMs, tree-based models, and reservoir computing, to deep-learning models [100], such as feedforward, convolutional, or recurrent neural networks. Here, we will exclusively consider shallow models, with a strong emphasis on kernel machines, with fundamental concepts and results introduced below. These notions will be particularly useful in the understanding of the photonic kernel machine discussed in Sec. III and the associated physical-implementation model presented in Sec. IV.

### A. Shallow models

The principle underlying the learning process of shallow models is sketched in Fig. 1. Inputs are first embedded from their original *input space* into a typically higher-dimensional *feature space* where the optimization problem becomes linear. The optimization process is then realized therein, for instance by means of standard convex optimization algorithms, and often reduces to the geometrical problem of identifying an optimal hyperplane.

More formally, the trial function $\hat{f}$ can be expressed in terms of a set of transformed (feature-space) coordinates as

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} w_m \tilde{x}_m + b = \mathbf{w}^T \tilde{\mathbf{x}} + b, \tag{1}$$

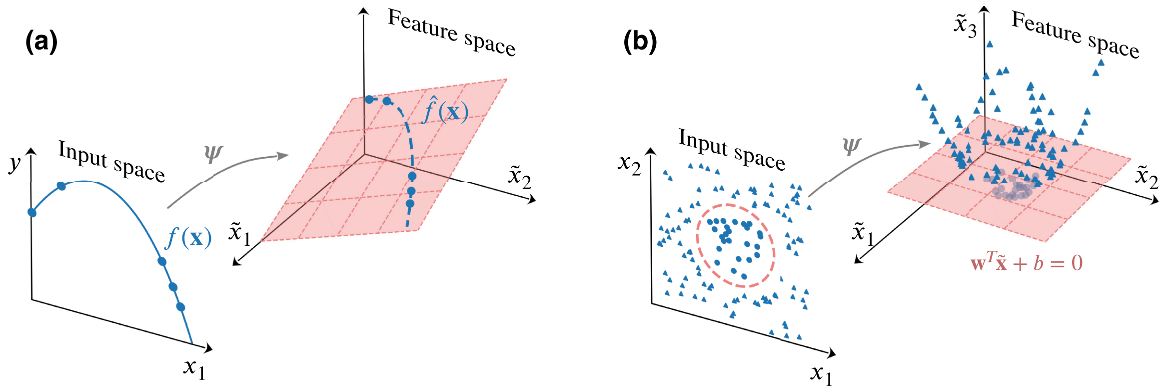



FIG. 1. Schematic representation of the embedding from input to feature space by a shallow model. (a) Example of regression: the quadratic function $f(x) = 3x - x^2 + 4$ is linearly fitted after the embedding of Eq. (1), with $\psi_1(x) = x^2$ and $\psi_2(x) = x$. Note that, in feature space, all data points lie on the same plane. (b) Example of binary classification: triangles and dots become linearly separable in feature space after the embedding of Eq. (1), with $\psi_1(\mathbf{x}) = x_1$, $\psi_2(\mathbf{x}) = x_2$, and $\psi_3(\mathbf{x}) = x_1^2 + x_2^2$. The resulting input-space decision boundary $\hat{f}(\mathbf{x}) = 0$ is represented by a dashed line. Origins are shifted to improve legibility.

where

$$\tilde{\mathbf{x}} := \boldsymbol{\psi}(\mathbf{x}), \tag{2}$$

with orthogonal components $\langle \psi_m, \psi_n \rangle_p = \mathbb{E}_p[\psi_m(\mathbf{x})\psi_n(\mathbf{x})] \propto \delta_{m,n}$, and $\mathbb{E}_p[\psi_m(\mathbf{x})\psi_n(\mathbf{x})] := \int d\mathbf{x} p(\mathbf{x})\psi_m(\mathbf{x})\psi_n(\mathbf{x})$. Here the *feature map* $\boldsymbol{\psi} : \mathbf{x} \mapsto \tilde{\mathbf{x}}$ defines an embedding from the *input space* into a *feature space* of dimension $M \leq +\infty$, with a new associated set of coordinates $\tilde{\mathbf{x}} = [\tilde{x}_1, \ldots, \tilde{x}_M]^T$ [101]. Then the trial function (1) takes the form of the equation of a hyperplane of parameters $(\mathbf{w}, b)$ lying within this feature space.

As shown in Fig. 1(a), a nonlinear function in input space can indeed become linear when expressed in a higher-dimensional feature space spanned by nonlinear transformations of the inputs. The feature-space embedding maps the original data points onto a hyperplane therein. The model of Eq. (1) can thus perform regression by approximating the target function as a plane parametric curve $\tilde{\mathbf{x}} = \boldsymbol{\psi}(\mathbf{x})$ lying on this possibly infinite-dimensional hyperplane, whose parameters $(\mathbf{w}, b)$ are to be determined.

Binary classification can be operated in a similar way. Inputs $\mathbf{x}$ are associated to some class $a$ if $\hat{f}(\mathbf{x}) > 0$ and to the complementary class $b$ otherwise. While this function might be involuted in input space, it may also become linear in some suitable feature space, the frontier between the two classes, $\hat{f}(\mathbf{x}) = \mathbf{w}^T\tilde{\mathbf{x}} + b = 0$, thus becoming the equation of a plane. In this situation, the objective of the optimization becomes finding the proper plane in feature space that separates data belonging to the two distinct classes into two separate clusters, as illustrated in Fig. 1(b). This linear decision boundary translates back into a potentially nontrivial one in input space [dashed circle in Fig. 1(b)]. The same mechanism can be exploited to perform $k$-class classification via the one-versus-one and one-versus-the-rest techniques, which split the problem into respectively $k(k-1)/2$ and $k$ binary-classification problems.

Such a shallow architecture presents a number of advantages. First, the optimization process is most often convex, in many cases analytic. Second, because it relies on a fixed feature-space embedding process, independent of any parameter to be updated during the training process, this architecture can be implemented on fast physical hardware, thus lifting any further analog-to-digital interfacing overhead.

Many possible choices exist to realize the above feature-space embedding. One popular choice is that of reservoir computing [29,30]. This scheme exploits the nonlinear response of a dynamical system with many degrees of freedom (the *reservoir*) to a driving input that encodes the data to be analyzed. As the set of produced independent nonlinear outputs for a single input is increased, the spanned feature space is expected to "include" the optimal one where the problem becomes linear [20,21]. While no control on the encoding of the inputs and the parameters of the reservoir is in principle required, in practice this may impact the convergence of the performance of the model in the number of extracted nonlinear outputs.

In the following, a similar alternative approach will be considered, rooted in the theory of SVMs: the so-called *kernel machines*. These rely on an educated guess: the feature-space embedding is induced based upon a choice of similarity metric, encoded in a *kernel* $K$, which associates to any two given inputs $\mathbf{x}$ and $\mathbf{x}'$ a measure of their similarity $K(\mathbf{x}, \mathbf{x}')$. Beyond providing heuristics on the way of building a feature-space embedding best suited for a specific task, such models come with theoretical bounds on the probability of generalization error [98,102,103].

In practice, kernel-method practitioners choose a kernel suitable for their applications; among popular choices are linear kernels, $K(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{x}'$; polynomial kernels, $K(\mathbf{x},\mathbf{x}') = (1+\mathbf{x}^T\mathbf{x}'/c)^d$; radial basis functions, $K(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|_2^2/\sigma^2)$ [104,105]; and sigmoid kernels, $K(\mathbf{x},\mathbf{x}') = \tanh(\kappa\mathbf{x}^T\mathbf{x}'+\theta)$, to name a few [106]. Very recently, "quantum" kernels of the form $K(\mathbf{x},\mathbf{x}') = \mathrm{Tr}[\hat{\rho}(\mathbf{x})^\dagger\hat{\rho}(\mathbf{x}')]$ or $K(\mathbf{x},\mathbf{x}') = \langle\hat{A}(\mathbf{x})^\dagger\hat{A}(\mathbf{x}')\rangle$, where $\hat{\rho}(\mathbf{x})$ and $\hat{A}(\mathbf{x})$ denote quantum operators evolved through some input-dependent unitary transformation $\hat{U}[\mathbf{x}]$, were proposed [107–109] and even experimentally implemented [109–111]. Quantum advantage on a classification task was demonstrated on such kernel machines [112]; more recently, it was shown that circuit-based "quantum neural networks" could be described as such quantum kernel machines [113].

Although in principle large-scale data processing with kernels can be computationally very expensive, several approximate techniques have been developed to overcome this limitation [114–117], some of which have recently inspired physical implementations capable of efficiently performing approximate kernel evaluations of high-dimensional vectorial data (e.g., images) on optical hardware [11,118]; see also Ref. [34] for interesting related strategies using free-space photonics. In the present work, we both define a kernel suitable for processing (infinite-dimensional) continuous time signals and provide a way to efficiently approximate it with photonic hardware.

### B. Kernel machines

We here recall some fundamental identities from the theory of kernel machines that will prove useful in what follows.

Upon making a specific choice of symmetric positive semidefinite kernel $K$, the kernel machine's predictions $\hat{f}$ can be evaluated in two different ways. First, as an average over a set of trainable model parameters $\boldsymbol{\alpha} = [\alpha_1,\alpha_2,\ldots,\alpha_N]^T$, weighted over the similarity between the considered input $\mathbf{x}$ and all those composing the training set,

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N}\alpha_i K(\mathbf{x},\mathbf{x}^{(i)}) + b, \tag{3}$$

where $b$ is a bias parameter. Alternatively, predictions can be obtained as an expansion over some set of $M \le +\infty$ possibly nonlinear functions $\{h_m\}_{m=1}^M$ such that $K(\mathbf{x},\mathbf{x}') = \sum_m h_m(\mathbf{x})h_m(\mathbf{x}')$, for all $(\mathbf{x},\mathbf{x}')$. Explicitly,

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M}\beta_m h_m(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{h}(\mathbf{x}), \tag{4}$$

where $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_M]^T$ is the set of parameters of the model [119].

The optimal parameters of the model ($\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, respectively) are determined by minimizing a cost function. Each approach leads to a different yet equivalent optimization problem, respectively known as the *dual* and the *primal* problem, of the form

$$J(\boldsymbol{\xi}) = \sum_{i=1}^{N} V\big(y^{(i)},\hat{f}(\mathbf{x}^{(i)})\big) + \lambda R(\boldsymbol{\xi}), \tag{5}$$

with either $\boldsymbol{\xi} = \boldsymbol{\alpha}$ (dual problem) or $\boldsymbol{\xi} = \boldsymbol{\beta}$ (primal problem). Here $V(\mathbf{y}^{(i)},\hat{f}(\mathbf{x}^{(i)}))$ is some pointwise error function on the predictions made by the model that one wants to minimize and $R$ is a regularization term whose strength is controlled by means of a "bias" hyperparameter $\lambda$.

Regularization prevents the trained model from overfitting the dataset, thus increasing its generalization power. In the following, ridge regularization will be considered, defined by

$$R(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}, \quad R(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2; \tag{6}$$

where $K_{ij} = K(\mathbf{x}^{(i)},\mathbf{x}^{(j)})$ denotes the *kernel matrix*.

Both the dual and the primal approaches can be shown to lead to the same feature-space embedding, of the form of Eq. (1), with feature maps given by $\mathbf{x} \mapsto \psi_m(\mathbf{x}) = \sqrt{\gamma_m}\phi_m(\mathbf{x})$, where $\gamma_m$ and $\phi_m$ simply derive from the kernel's eigendecomposition:

$$K(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{M}\gamma_m\phi_m(\mathbf{x})\phi_m(\mathbf{x}'), \tag{7}$$

with $\gamma_{m+1} \le \gamma_m$ and $\langle\phi_m,\phi_n\rangle_p = \mathbb{E}_p[\phi_m(\mathbf{x})\phi_n(\mathbf{x})] \equiv \int d\mathbf{x}\, p(\mathbf{x})\psi_m(\mathbf{x})\psi_n(\mathbf{x}) = \delta_{m,n}$. This eigendecomposition can be given empirical estimates by approximating the actual probability density distribution of inputs $p(\mathbf{x})$ by the *empirical* one $\hat{p}(\mathbf{x})$ [114], such that $\int d\mathbf{x}\hat{p}(\mathbf{x})f(\mathbf{x}) = (1/N)\sum_i f(\mathbf{x}^{(i)})$. As shown in the Appendix, given a primal problem parametrized by $\{h_m\}_{m=1}^M$, the *empirical eigenvalues* $\{\hat{\gamma}_m\}_{m=1}^{\mathrm{rank}(\mathbf{K})\le M}$ are simply given by those of the matrix $\mathbf{k}/N = \mathbf{H}^T\mathbf{H}/N$, where $H_{im} = h_m(\mathbf{x}^{(i)})$, with the following associated *empirical eigenfunctions*:

$$\hat{\phi}_m(\mathbf{x}) = \frac{1}{\sqrt{N\hat{\gamma}_m}}\mathbf{u}_m^T\mathbf{h}(\mathbf{x}), \tag{8}$$

where $\mathbf{u}_m$ denotes the $m$th eigenvector of $\mathbf{k}$. Together these lead to the *empirical feature map* $\hat{\psi}_m(\mathbf{x}) = \sqrt{\hat{\gamma}_m}\hat{\phi}_m(\mathbf{x})$.

The ability of kernel machines to filter out noise in input space can now be understood in very simple terms. To this end, let us start from the primal approach and consider a kernel machine whose predictions are obtained as in Eq. (4), with outputs centered by the

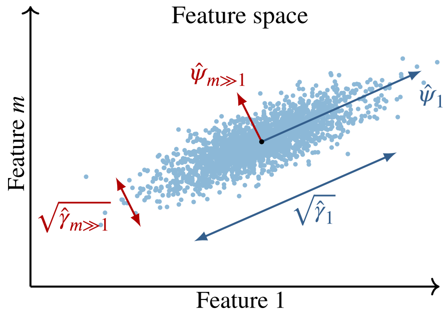



FIG. 2. Schematic representation of the projection of the feature-space-embedded data onto the plane spanned by the leading principal component of some kernel $K$ and some subleading one. The variance stemming from the correlations of the embedded training examples is expected to be mostly contained within the dimensions spanned by the kernel's principal components, whereas the variance of the noise is expected to be contained along the remaining (orthogonal) directions.

empirical mean over the training set, that is, replacing $h_m(\mathbf{x})$ by $\tilde{h}_m(\mathbf{x}) = h_m(\mathbf{x}) - \mathbb{E}_{\hat{p}}[h_m(\mathbf{x})]$, where $\mathbb{E}_{\hat{p}}[X(\mathbf{x})] = (1/N)\sum_i X(\mathbf{x}^{(i)})$, for any $X$. It then follows that the corresponding kernel's empirical eigenfunctions are centered as well, $(1/N)\sum_i \hat{\phi}_m(\mathbf{x}^{(i)}) = 0$, and the eigendecomposition (7) is in all respects analogous to principal component analysis on the feature-space-embedded data inputs [98,120,121]. In this picture, schematically illustrated in Fig. 2, the empirical eigenvalues exactly correspond to the variance of the data along every feature-space direction:

$$\hat{\gamma}_m = \mathbb{E}_{\hat{p}}\left[\left(\hat{\psi}_m(\mathbf{x}) - \mathbb{E}_{\hat{p}}[\hat{\psi}_m(\mathbf{x})]\right)^2\right]. \tag{9}$$

Large variances of this type correspond to highly correlated features that maximally deviate between inputs considered as "dissimilar" according to one's initial choice of similarity metric. Conversely, low variances correspond either to redundant features or to uncorrelated noise contained in the inputs.

By now decomposing the trial function in the kernel basis identified above,

$$\hat{f}(\mathbf{x}) = \sum_m \langle \hat{\phi}_m, \hat{f}\rangle_{\hat{p}} \hat{\phi}_m(\mathbf{x}) \equiv \sum_m \hat{f}_m \hat{\phi}_m(\mathbf{x}), \tag{10}$$

where the components $\hat{f}_m$ may be interpreted as degrees of freedom of the model, it follows from Eq. (6) that the regularization penalty can be rewritten as

$$\lambda R(\boldsymbol{\alpha}) = \sum_m \frac{\lambda_m}{2}|\hat{f}_m|^2, \quad \lambda_m = \frac{\lambda}{\hat{\gamma}_m}. \tag{11}$$

This means that the penalty on the magnitude of the $m$th component of $\hat{f}$ during the optimization is inversely proportional to the variance of the associated feature-space coordinate $\psi_m(\mathbf{x})$. Therefore, the effect of the regularization on the optimization is to impose a soft cutoff on the dimensionality of the feature-space hyperplane schematically illustrated in Fig. 1, while preserving most of the original variance contained in the inputs by only keeping the most statistically relevant directions, expected to encode the correlated information of interest in the inputs, while filtering out the noise [121]. This can be understood as a filtering process, as will clearly appear in what follows.

## III. PHOTONIC KERNEL MACHINES

The spectral analysis of rf signals is a very broad field with countless applications. We notably focus on the analysis of ultrashort pulsed signals. This is relevant, for instance, in the field of pulse-Doppler radars. Such devices emit a pulsed rf signal around a carrier frequency with a given repetition rate, and receive it back via an antenna after it has been reflected by a moving target. This reflected signal is measured and numerically processed to extract information about the position and the velocity of the tracked target. The repetition rate is limited by the processing time, which is in turn bounded by the acquisition time. Nowadays, state-of-the-art performance is obtained by direct sampling of the signal sensed by the antenna. Yet, the sampling rate of state-of-the-art rf analog-to-digital converters lies typically below 10 gigasamples per second [122]. In order to measure 1000 samples of a pulse of interest with such a converter, one needs the pulse length to exceed 25 $\mu$s, limiting our ability to analyze shorter pulses. This issue may be circumvented by operating several such samplers in parallel. However, the resulting devices are very expensive, heavy, and bulky.

This context motivates the quest for technologies that go beyond these limitations. In what follows, we propose to resort to a learning all-optical processing device based upon the kernel-machine concept. We will first introduce a similarity kernel ideally suited for such spectral analysis tasks. We will show that this kernel can be evaluated rather naturally by means of an optical single-shot intensity measurement at the output of a photonic lattice. Such a photonic kernel machine makes it possible to process complex rf signals in times of the order of tens or hundreds of picoseconds.

### A. Theoretical description

As we have seen above, learning with kernels starts from the introduction of a kernel that serves as a measure of similarity between inputs. In particular, we focus the discussion here on the spectral analysis of ultrashort pulsed rf signals. The spectral information of such signals $s(t)$ is encoded in their energy spectral density $S[\omega] = |s[\omega]|^2$, where $s[\omega] = (2\pi)^{-1/2}\int dt e^{-i\omega t}s(t)$. Rather naturally, the similarity between two signals $s(t)$ and $s'(t)$ can thus be given a general expression of the form

$$K(S, S') = \int d\omega d\omega' S[\omega]\mathcal{K}(\omega - \omega')S'[\omega'], \tag{12}$$

where $\mathcal{K}(\omega - \omega')$ is a function peaking around $\omega = \omega'$ and with a typical width $\delta\omega$. Such a kernel compares two given input signals by contrasting their energy spectral densities at each frequency with a certain tolerance on their fine structure at frequency scales below $\delta\omega$. While this provides a good and flexible similarity metric, the numerical evaluation of such a kernel is rather unsuitable in practice using current electronics. Indeed, this would require several costly steps: (i) each pulse would have to be sampled in time at similar sampling rates over a time interval longer than $2\pi/\delta\omega$; (ii) the digitized signals would then have to be numerically Fourier-transformed and stored; (iii) finally, each kernel evaluation would involve calculating a double integral. In contrast, we will see that the feature-space embedding of this kernel can be performed rather naturally on photonic hardware. For that purpose, the rf signal to be analyzed can first be imprinted on an optical carrier injected into the hardware.

Let us consider an optical system consisting of a set of $M$ generic normal modes as described by the set of bosonic operators $\{\hat{\alpha}_\ell\}_{\ell=1}^M$. Their response to any $i$th pulsed input signal $s^{(i)}(t)$ of interest is well described by a quantum Langevin equation [123,124]. In frequency space, this takes the form

$$\hat{\alpha}_\ell[\omega] = \chi_\ell[\omega]\Big(is^{(i)}[\omega] + \sqrt{\kappa/2}\hat{\alpha}_\ell^{\text{in}}[\omega]\Big), \tag{13}$$

where $\hat{\alpha}_\ell[\omega] = (2\pi)^{-1/2}\int dte^{-i\omega t}\hat{\alpha}_\ell(t)$ is the Fourier-transformed annihilation operator associated to the $\ell$th normal mode, $\chi_\ell[\omega]$ its (model-dependent) susceptibility at frequency $\omega$, and $\kappa$ is the loss rate, taken independent of $\ell$. Here, the Langevin input fields $\hat{\alpha}_\ell^{\text{in}}(t)$ account for zero-temperature quantum noise and satisfy the usual expectation values: $\langle\hat{\alpha}_\ell^{\text{in}}(t)\rangle = \langle\hat{\alpha}_\ell^{\text{in}\dagger}(t)\hat{\alpha}_{\ell'}^{\text{in}}(t')\rangle = 0$ and $\langle[\hat{\alpha}_\ell^{\text{in}}(t), \hat{\alpha}_{\ell'}^{\text{in}\dagger}(t')]\rangle = \delta_{\ell,\ell'}\delta(t - t')$.

Optical populations in the modes, induced by the input pulse, are measured via the radiated optical power collected by a detector, with some integration time $\Delta t$ much larger than the length of the pulse, $\Delta t \gg 2\pi/\Delta\omega$, where $\Delta\omega$ is the bandwidth of the signal to be analyzed. In this limit, the measured normal-mode populations are expressed as

$$\begin{aligned}\bar{n}_\ell^{(i)} &\simeq \frac{1}{\Delta t}\int dt\langle\hat{\alpha}_\ell^\dagger(t)\hat{\alpha}_\ell(t)\rangle = \frac{1}{\Delta t}\int d\omega\langle\hat{\alpha}_\ell^\dagger[\omega]\hat{\alpha}_\ell[\omega]\rangle \\ &= \int d\omega h_\ell(\omega)S^{(i)}[\omega] \equiv \langle h_\ell|S^{(i)}\rangle,\end{aligned} \tag{14}$$

where the integration window was approximately extended to $\mathbb{R}$ in the first equality, $h_\ell(\omega) = \Delta t^{-1}|\chi_\ell[\omega]|^2$ is the optical population susceptibility, and $S^{(i)}[\omega] = |s^{(i)}[\omega]|^2$ the energy spectral density of the $i$th pulse.

From such a population measurement, vector-valued predictions on any $i$th input energy spectral density $S$ take the form of those of a kernel machine as in Eq. (4):

$$\hat{f}(S) = \mathbf{B}^T\bar{\mathbf{n}}(S), \tag{15}$$

with components

$$\hat{f}_n(S) = \sum_\ell B_{\ell n}\langle h_\ell|S\rangle, \tag{16}$$

where $\mathbf{B}$ is a matrix of parameters, $S$ plays the role of the input, and the feature index $m$ in Eq. (4) is replaced by that of the optical modes $\ell$. For optical relaxation times $(2\pi\kappa^{-1})$ of the order of tens of picoseconds, such predictions can be realized for ultrashort rf pulses at a throughput in excess of tens of gigahertz.

Let us check that the corresponding dual-picture kernel is indeed of the form of Eq. (12) under some general assumptions on the optical-mode density spectrum of the photonic system. The kernel we consider takes the form

$$\begin{aligned}K(S, S') = \bar{\mathbf{n}}^T(S)\bar{\mathbf{n}}(S') &\equiv \sum_\ell\langle S|h_\ell\rangle\langle h_\ell|S'\rangle \\ &= \int d\omega d\omega' S[\omega]\mathcal{K}(\omega, \omega')S'[\omega'],\end{aligned} \tag{17}$$

with $\mathcal{K}(\omega, \omega') := \sum_\ell h_\ell(\omega)h_\ell(\omega')$ an integral kernel [125]. Provided all normal modes share the same susceptibility, $h_\ell(\omega) = h(\omega - \omega_\ell)$, and for a continuous optical-mode density spectrum $\rho(\omega)$, this amounts to

$$\mathcal{K}(\omega, \omega') = \int d\Omega\rho(\Omega)h(\omega - \Omega)h(\omega' - \Omega). \tag{18}$$

By further assuming a smooth spectral density of width larger than the normal-mode linewidth, one has

$$\mathcal{K}(\omega, \omega') \simeq \rho\left(\tfrac{\omega+\omega'}{2}\right)\int dth(t)h(-t)e^{-i|\omega-\omega'|t}. \tag{19}$$

The integral kernel is thus completely determined by the observable susceptibility, here the mode population via $h(\omega) = |\chi[\omega]|^2/\Delta t$. For instance, the susceptibility of a linear optical mode $\chi_\ell[\omega] = 1/(-i(\omega - \omega_\ell) + \kappa/2)$ translates into a Lorentzian,

$$\mathcal{K}(\omega, \omega') \propto \frac{1}{(\omega - \omega')^2 + \kappa^2}, \tag{20}$$

shown to perform better on some tasks than the more common radial basis function [126]. This integral kernel indeed takes the desired form introduced in Eq. (12), with a spectral tolerance $\delta\omega \sim \kappa$ given by the mode's linewidth.

### B. Inspecting the feature space

One can access to the feature space associated to such a photonic kernel machine as well. Indeed, the above kernel admits an eigendecomposition analogous to Eq. (7),

$$K(S,S') = \sum_\ell \langle S|h_\ell\rangle\langle h_\ell|S'\rangle = \sum_\ell \gamma_\ell \langle S|\phi_\ell\rangle\langle\phi_\ell|S'\rangle\,, \quad (21)$$

with empirical eigenfunctionals and feature maps as given by Eq. (8),

$$\hat{\phi}_\ell(\omega) = \frac{1}{\sqrt{N\hat{\gamma}_\ell}}\mathbf{u}_\ell^T\mathbf{h}(\omega), \quad \hat{\psi}_\ell(\omega) = \sqrt{\hat{\gamma}_\ell}\hat{\phi}_\ell(\omega), \quad (22)$$

where $\mathbf{u}_\ell$ and the empirical eigenvalues $\hat{\gamma}_\ell$ are obtained following the previous section as the $\ell$th eigenvector and eigenvalue of the matrix $\mathbf{k}/N = \mathbf{H}^T\mathbf{H}/N$, with now $H_{i\ell} = \langle h_\ell|S^{(i)}\rangle = \bar{n}_\ell^{(i)}$. Strikingly, this kernel matrix, and thus the above feature maps, can be directly constructed in experiments from the optical-population measurements over the training set as

$$\frac{1}{N}k_{\ell\ell'} = \frac{1}{N}\sum_i \bar{n}_\ell^{(i)}\bar{n}_{\ell'}^{(i)}. \quad (23)$$

Now that the feature map is identified, let us examine how the geometric picture introduced in Sec. II A translates to the photonic kernel case. In particular, we consider a regression problem where the targeted quantity is a function of the angular frequency $y[\omega]$, as is the case, for instance, when trying to estimate the spectrum of a pulse ($y[\omega] = S[\omega]$). This scenario can be dealt with by frequency-binning the target as $y_n = y[\omega_n]$ and trying to find its best fit using an estimator of the form $\hat{y}^{(i)}[\omega_n] := \hat{f}_n(S^{(i)})$, for any pulse $i$. Then it follows from Eqs. (16) and (22) that predictions can be rewritten in the form of Eq. (1), as an expansion over feature-space coordinates:

$$\hat{y}^{(i)}[\omega_n] = \sum_\ell \hat{w}_\ell[\omega_n]\langle\hat{\psi}_\ell|S^{(i)}\rangle, \quad (24)$$

where

$$\hat{w}_\ell[\omega_n] = \sqrt{N}[\mathbf{u}^T\hat{\mathbf{B}}]_{\ell n}, \quad (25)$$

with $\mathbf{u} = [\mathbf{u}_1,\ldots,\mathbf{u}_M]$. Here, the quantities $\tilde{S}_\ell \equiv \langle\hat{\psi}_\ell|S\rangle$ play the role of feature-space coordinates [$\tilde{x}_m$ in Eq. (1)] and can be interpreted as reciprocal-space components of the input $S[\omega]$ with respect to the basis $\{\hat{\psi}_\ell\}_\ell$. The quantities $\hat{w}_\ell[\omega_n]$ correspond to the parameters of the feature-space hyperplane [$w_m$ in Eq. (1)] and can be interpreted as a set of "functions" $\{\omega \mapsto \hat{w}_\ell[\omega]\}_\ell$ learned during the optimization procedure. These may be evaluated explicitly at the chosen frequency bins $\{\omega_n\}_n$ from the measurement data and the trained parameters $\hat{\mathbf{B}}$ by making use of the above expression. This picture is schematically illustrated in Fig. 3, in analogy with Fig. 1(a), on the spectral-analysis task.

Let us note that by setting the trainable parameters to $\hat{w}_\ell[\omega_n] = \hat{\psi}_\ell(\omega_n)/\hat{\gamma}_n$ in the absence of any optimization step, one has directly that $\hat{y}^{(i)}[\omega_n] = S^{(i)}[\omega_n]$. This means that a photonic kernel machine is at least able to reproduce the energy spectral density of a pulse $s(t)$ from a single-shot intensity measurement, provided the energy spectral density belongs to the linear span of $\{\hat{\psi}_\ell\}_\ell$, as will indeed be verified in Sec. V. The action of the ridge regularization of Eq. (6) on such a spectral-analysis task can be analytically uncovered. Indeed, upon choosing mean square error as error function, $V(y,\hat{y}) = \sum_n|y[\omega_n] - \hat{y}[\omega_n]|^2$, the functions learned by the model can be shown to be given by

$$\hat{w}_\ell[\omega]\Big|_\lambda = \frac{1}{1+\lambda/\hat{\gamma}_\ell} \times \hat{w}_\ell[\omega]\Big|_{\lambda=0}, \quad (26)$$

and are thus filtered by the photonic kernel machine in accordance with the inverse of the variance of their

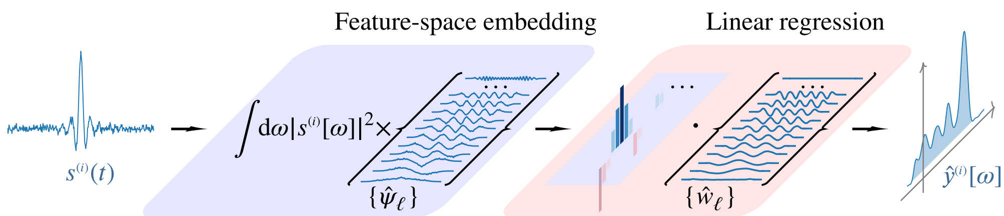



FIG. 3. Illustration of the photonic-kernel-machine processing mechanism. The measurement of the pulse-induced optical populations embeds the pulse's energy spectral density into some reciprocal space by projection onto a set of orthogonal base functions $\{\hat{\psi}_\ell\}_\ell$. The spectrum of the $i$th pulse is reconstructed by a linear combination $\hat{y}^{(i)}[\omega] = \sum_\ell w_\ell[\omega]\langle\hat{\psi}_\ell|S^{(i)}\rangle$ of its reciprocal-space components (features) over some set of learned functions $\{\hat{w}_\ell\}_\ell$. In this example, the learned functions $\{\hat{w}_\ell\}_\ell$ are filtered analogs of $\{\hat{\psi}_\ell\}_\ell$ and the photonic kernel machine is able to extract the spectrum of the incoming pulse from its noisy background.

associated features over the training set:

$$\hat{\gamma}_\ell = \mathbb{E}_{\hat{p}}\Big[\big(\langle S|\hat{\psi}_\ell\rangle - \mathbb{E}_{\hat{p}}[\langle S|\hat{\psi}_\ell\rangle]\big)^2\Big]. \tag{27}$$

## IV. PHYSICAL IMPLEMENTATION

We apply the above theory to the ultrafast processing of rf pulsed signals with a photonic lattice. The task involves analyzing a set of baseband rf signals $\{s^{(i)}(t)\}$ over a bandwidth $\Delta\omega$ around some reference angular frequency $\omega_0$ by extracting some associated quantity of interest $\mathbf{y}$, such as the energy spectrum or the peak frequency. To this end, let us introduce an adapted physical implementation of a photonic-lattice-based kernel machine. The setup is schematically illustrated in Fig. 4. It involves four successive elements.

Input signals enter the first unit of the system via an electro-optic modulator through frequency mixing with an optical carrier, $c(t) = c_0 \exp(-i\omega_p t)$, resulting in a modulated signal of the form $F^{(i)}(t) = s^{(i)}(t)e^{-i\omega_p t}$ whose central angular frequency $\omega_p + \omega_0$ may be shifted to accommodate the processing of signals in very different bands. The angular frequency of the local oscillator is set to $\omega_p = \bar{\omega} - \omega_0$, where $\bar{\omega}$ denotes the central angular frequency of the lattice's normal modes, to maximize the response of the system. The now modulated optical signal is then routed to the photonic lattice, entering the cavities as a coherent drive.

The second unit consists of an $L \times L$ quadratic photonic lattice whose cavities are mutually coupled via near-field nearest-neighbor interactions. These cavities are coupled to the external modulated drive with some arbitrarily spatially dependent weight $v_\ell$. In a frame rotating at the drive frequency, the dynamics of such a lattice is described by the following set of quantum Langevin equations [123,124]:

$$\partial_t \hat{a}_\ell(t) = \big[i(\Delta_\ell + \omega_0) - \kappa_\ell/2\big]\hat{a}_\ell(t) + i \sum_{\ell' \in N(\ell)} J_{\langle \ell',\ell\rangle}\hat{a}_{\ell'}(t) + i v_\ell s(t) + \sqrt{\kappa_\ell/2}\, \hat{a}_\ell^{\rm in}(t), \tag{28}$$

where $\hat{a}_\ell$ is the photon annihilation operator for site $\ell$, $\Delta_\ell = \omega_p - \omega_\ell$ is the detuning of the local oscillator, $\kappa_\ell$ the optical relaxation rate, $J_{\langle \ell',\ell\rangle}$ the linear coupling rate between cavities $\ell$ and $\ell'$, $s(t)$ the broadband driving signal, and $v_\ell$ the local weight of the coupling of the $\ell$th cavity to the external drive. Operators $\hat{a}_\ell^{\rm in}(t)$ account for quantum noise and satisfy $\langle \hat{a}_\ell^{\rm in}(t)\rangle = \langle \hat{a}_\ell^{\rm in\dagger}(t)\hat{a}_{\ell'}^{\rm in}(t')\rangle = 0$ and $\langle[\hat{a}_\ell^{\rm in}(t), \hat{a}_{\ell'}^{\rm in\dagger}(t')]\rangle = \delta_{\ell,\ell'}\delta(t-t')$. A high degree of control of the parameters is not required. As a matter of fact, in the following numerical simulations, while angular frequencies $\omega_\ell$ will be uniformly set for simplicity, with $\Delta = -\omega_0$, the remaining parameters will be random variables. In particular, $J_{\langle \ell',\ell\rangle}$ will be uniformly drawn in the interval $[0, J_{\max}]$, $\kappa_\ell$ normally distributed around $\bar{\kappa} = zJ_{\max}/40$

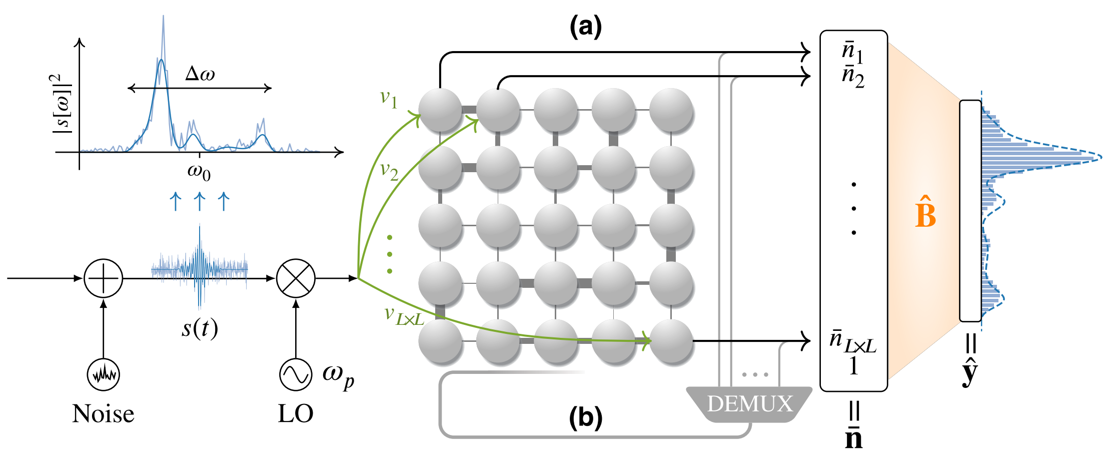



FIG. 4. Schematic representation of a trained photonic-lattice-based kernel machine estimating the energy spectral density of a noisy pulse from single-shot intensity measurements. The processing mechanism is as follows. (i) A pulsed rf signal $s^{(i)}(t)$ is sent to the device in the presence of white noise. (ii) The noisy signal is then transferred to an optical carrier of angular frequency $\omega_p$ thanks to an electro-optic modulator ($\otimes$) and (iii) coherently injected into a lattice of $L \times L$ linear cavities with random parameters. (iv) The resulting optical populations are then measured and, finally, (v) the spectrum of the original noiseless signal is reconstructed by linearly combining the measured quantities thanks to the fixed matrix $\hat{\mathbf{B}}$, obtained from the previous training protocol. (a) Local-mode measurement protocol: populations are measured through the light locally radiated by the cavities. (b) Normal-mode measurement protocol: the cavities' optical field is first collected by an evanescently coupled waveguide and frequency-demultiplexed into $L \times L$ channels, whose optical populations are finally measured. Indeed, each normal mode has a different central frequency, allowing the corresponding populations to be spectrally separated.

($z = 4$) with a standard deviation of 10%, $\Delta\omega = zJ_{\max}$, and $\mathbf{v}$ will be set to a normalized random real vector.

The third unit consists of a sensor that measures the intensity of the light radiated by the optical cavities as they are externally driven by an input signal. These populations are measured by time-averaging over some detector integration time $\Delta t$ and collected into a vector $\bar{\mathbf{n}}^{(i)} = [\bar{n}_1^{(i)}, \ldots, \bar{n}_{L\times L}^{(i)}, 1]^T$, where a unit entry is added to get a supplementary trainable parameter acting as a bias. In what follows, two alternative measurement settings will be considered.

(i) Measure of the local populations: $\bar{n}_\ell^{(i)} = (1/\Delta t)\int dt\, \langle \hat{a}_\ell^\dagger(t)\hat{a}_\ell(t)\rangle$. This corresponds, for instance, to the experimental situation where the intensity that is vertically emitted by the cavities is measured by a camera facing the lattice. This is schematically illustrated in Fig. 4(a).

(ii) Measure of the normal-mode populations: $\bar{n}_\ell^{(i)} = (1/\Delta t)\int dt \langle \hat{\alpha}_\ell^\dagger(t)\hat{\alpha}_\ell(t)\rangle$, where $\hat{\alpha}_\ell$ denotes the annihilation operator on the $\ell$th normal mode. This corresponds, for example, to the experimental situation where the field leaking from the cavities is collected by an evanescently coupled waveguide and frequency-demultiplexed into $L \times L$ frequency channels coupled to photodetectors. This is schematically illustrated in Fig. 4(b).

Finally, a fourth unit performs a linear combination of the measured populations by acting with a $n \times (L^2 + 1)$ matrix of parameters $\mathbf{B}$, to be optimized by training. The output of this last unit is thus of the form $\hat{\mathbf{y}}^{(i)} = \mathbf{B}^T\bar{\mathbf{n}}^{(i)}$, as in Eq. (15).

### A. Training the lattice-based photonic kernel

The training starts from a training and a testing set composed of $N_{\text{train}}$ and $N_{\text{test}}$ signals, respectively, each of the form $\{(s^{(i)}, \mathbf{y}^{(i)})\}_i$, where $s^{(i)}$ denotes the $i$th signal and $\mathbf{y}^{(i)}$ a set of known associated features we want our system to learn how to estimate. In regression tasks, these labels $\mathbf{y}^{(i)}$ may take arbitrary values, whereas in binary-classification tasks $y^{(i)} = \pm 1$ depending on whether the associated input $s^{(i)}$ belongs to a targeted class or not.

The trial function of the untrained model is initially given by $\hat{f}(s^{(i)}) = \mathbf{B}^T\bar{\mathbf{n}}^{(i)}$ [127]. The training is carried out as follows.

(i) *Construct the matrix* $\mathbf{Y}$. From the known features $\{\mathbf{y}^{(i)}\}_i$ associated to the samples of the training set, $\mathbf{Y}$ is first computed as $Y_{im} = y_m^{(i)}$.

(ii) *Obtain the matrix* $\mathbf{H}$. Each input signal $s^{(i)}$ of the training set is fed into the device at the electro-optic modulator and the resulting cavity populations $\bar{\mathbf{n}}^{(i)}$ are measured by the sensor. All these measured populations are stored in a matrix $H_{im} = \bar{n}_m^{(i)}$. Note that this evaluation is only to be performed once.

(iii) *Determine the optimal parameters* $\hat{\mathbf{B}}$. For a chosen value of the regularization hyperparameter $\lambda$, one obtains the optimal parameters by minimizing the cost function as $\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} J(\mathbf{B})$, with $J(\mathbf{B}) = \sum_{i=1}^{N_{\text{train}}} V(\mathbf{y}^{(i)}, \hat{f}(s^{(i)})) + \lambda/2\|\mathbf{B}\|_2^2$, where $\hat{f}(s^{(i)}) = \mathbf{B}^T\bar{\mathbf{n}}^{(i)}$, with $\bar{\mathbf{n}}^{(i)}$ as obtained in the previous step. The choice of loss function $V$ depends on the task. For regression tasks, a popular choice is simply $V(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, which corresponds to a least-squares problem; this directly yields $\hat{\mathbf{B}} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbb{1})^{-1}\mathbf{H}^T\mathbf{Y}$ analytically, with $\mathbf{Y}$ and $\mathbf{H}$ as computed at steps (i) and (ii), respectively. In binary classification ($y = \pm 1$), it is customary to choose a margin-maximizing loss functions [128]. Popular choices [99] are the hinge loss, $V(y, \hat{f}(s)) = \max[0, 1 - y\hat{f}(s)]$, which makes $\hat{f}$ directly approximate the class label $f(s) = y$, or the binomial deviance $\ln(1 + e^{-y\hat{f}(s)})$, which instead makes $\hat{f}$ approximate $f(s) = \ln[\mathbb{P}(y = +1|s)/\mathbb{P}(y = -1|s)]$. Upon choosing a loss function, the problem is convex and can be solved either analytically, for the square error, or by means of a convex optimization solver, using standard iterative methods.

(iv) *Evaluate the accuracy of the model*. Once the model is trained, for any new input signal $s$ fed into the system, features are estimated according to $\hat{f}(s) = \hat{\mathbf{B}}^T\bar{\mathbf{n}}$, from the measurement of the resulting populations $\bar{\mathbf{n}}$. Its accuracy can then be benchmarked on the testing set by comparing the predictions $\hat{\mathbf{y}}^{(j)} = \hat{f}(s^{(j)})$ against the known features $\mathbf{y}^{(j)}$, for every input signal $s^{(j)}$ in the testing set, to which the model was yet never exposed. This is done through a metric that may differ from the cost function, in particular if regularization was employed.

### B. Benchmarking the photonic-lattice kernel machine

In what follows, two different approaches will be employed to benchmark the above-defined physical implementation of a photonic kernel.

In order to evaluate the ability of the model to estimate the spectrum of pulsed signals, we will first use the squared modulus of the fast Fourier transform (FFT) of the input signals as a reference for energy spectral density. We will assume an ideal sampling of the pulse over a centered window of time length $\Delta T = 5 \times 2\pi/\kappa$ at a sampling rate of $f_s = 200 \times \kappa/2\pi$. For cavities with $2\pi/\kappa \sim 10$ ps, this corresponds to a sampling rate of $f_s = 10$ THz. Note that this is more than three orders of magnitude beyond the state of the art [119]. Yet in the following the photonic kernel will be shown to outperform this rather fictional ideal device.

We will also use a reservoir-computing approach based on a nonlinear polaritonic lattice as first introduced in Ref. [54]. This reservoir was numerically proven successful in image- and speech-recognition tasks and chaotic time-series forecasting, and was recently experimentally tested on an optical character-recognition task [55]. This reservoir is modeled by the following discrete complex Ginzburg-Landau equation:

$$\partial_t \alpha_\ell(t) = \left[ + i(\Delta_\ell + \omega_0) + \gamma - (\Gamma + ig)|\alpha_\ell(t)|^2 \right] \alpha_\ell(t) + i \sum_{\ell' \in N(\ell)} J_{\langle \ell', \ell \rangle} \alpha_{\ell'}(t) + i v_\ell s(t), \tag{29}$$

where $\gamma = P - \kappa/2$ is a gain coefficient that accounts for the single-body decay rate $\kappa/2$ and the magnitude of the external pumping $P$, which brings the system close to instability. Here $\Gamma$ and $g$ respectively account for two-body dissipation and interaction processes. The input signals are coherently injected in the same fashion as for the linear lattice. In the following numerical simulations, we will set $\Delta = -\omega_0$, $J_{\langle \ell', \ell \rangle}$ uniformly drawn in the interval $[0, J_{\max}]$, $\Gamma = zJ_{\max}/40$ $(z = 4)$, $\gamma/\Gamma = 8 \times 10^{-4}/2\pi$, $g/\Gamma = 1.6/2\pi$, $\Delta\omega = zJ_{\max}$, and $\mathbf{v}$ to a normalized random real vector. The output of this model depends on the amplitude of the input. In what follows, the input energy will be set to $50\Gamma$.

The training of this model is realized exactly as for the photonic kernel machine, from local population measurements $\bar{n}_\ell^{(i)} \propto (1/\Delta t) \int dt |\alpha_\ell(t)|^2$.

## V. APPLICATIONS

### A. Pulse spectral analysis

As a first illustration of the learning capabilities of the above-described photonic kernel machine, let us consider the extraction of the noiseless energy spectral density of an ultrashort pulsed rf signal embedded into a noisy background. Given an input noisy pulsed signal $s(t) + \xi(t)$, this task involves giving the best possible estimation $\hat{y}_n$ of its frequency-binned energy spectral density $y_n = |s[\omega_n]|^2$ regardless of the noisy background $\xi(t)$.

In order to train and study the performance of the model, we generate training and testing sets of $N_{\text{train}}$ and $N_{\text{test}}$ pulses, respectively. To do so, we first generate known noiseless random spectra of total energy $E_0 = \int d\omega S[\omega]$ by cubic B-spline interpolation of a set of $N_b$ signal bins $\tilde{s}_n := \tilde{s}[\omega_n]$, with frequency bins $\{\omega_n\}_n$ equally spaced within the band $[\omega_0 - \Delta\omega/2, \omega_0 + \Delta\omega/2]$ and randomly sampled from a Boltzmann distribution $p[\tilde{s}_n] = (1/Z)e^{-\beta V[\tilde{s}_n]}$ parametrized by the potential

$$V[\tilde{s}_n] = \frac{a}{N_b} \sum_{n=1}^{N_b - 1} |\tilde{s}_{n+1} - \tilde{s}_n|^2 + b \left| \max_n |\tilde{s}_n|^2 - S_{\text{peak}} \right|^2 + c \left| \operatorname{std}_n |\tilde{s}_n|^2 - \bar{S} \right|^2, \tag{30}$$

with $\tilde{s}_i = 0, i = 1, N_b$, $S_{\text{peak}} = 8\bar{S}$, $\bar{S} = E_0/\Delta\omega$. Here $a$ acts as stiffness parameter, whereas the $b$ term favors peaked spectra. Finally, the $c$ term prevents the sampled spectra from sharing too similar shapes by favoring the presence of secondary peaks. The parameters used throughout this section are $N_b = 20$, $\beta a = \beta c = 100$, and $\beta b = 50$. The energy spectral densities of the random noiseless spectra obtained are shown in Fig. 5(a), where the average energy distribution is compared to $\bar{S}$ and a few typical examples are plotted, exhibiting various peaks with different heights. The choice of bandwidth ($\Delta\omega = zJ_{\max}$) ensures that all the power of the signal to be analyzed can be sensed by the photonic lattice, as shown in Fig. 5(b), where the average optical spectrum of a large random photonic lattice is plotted. The random spectra are then Fourier-transformed to the time domain and white noise $\xi(t)$ is added to match some signal-to-noise ratio SNR, here defined as the relative contribution of the noise to the total energy in the analyzed band, $\text{SNR} = E_0^{-1} \int_{\omega_0 - \Delta\omega/2}^{\omega_0 + \Delta\omega/2} d\omega |\xi[\omega]|^2$.

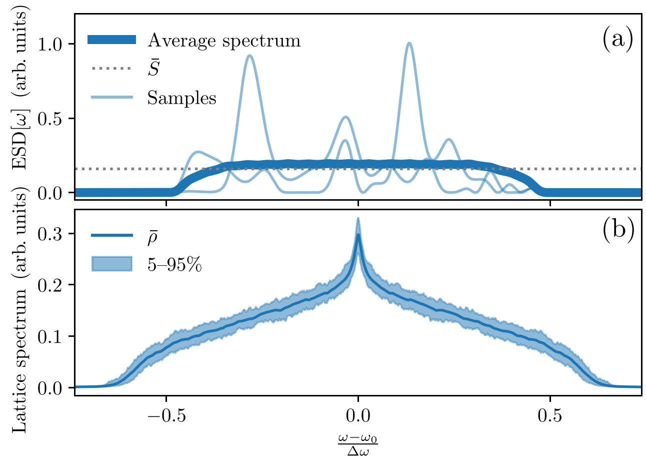



FIG. 5. (a) Energy spectral density averaged over $10^5$ realizations of the noiseless pulses, as given by Eq. (30), as well as that of a few individual realizations. (b) Mean and interval between 5th and 95th percentiles for the optical spectra of 100 realizations of a $30 \times 30$ random photonic lattice.

We then simulate the response of the photonic lattice to each of the driving input signals $s^{(i)}(t)$ of the training set by numerically integrating the coupled dynamical equation (28). For each signal, either local or normal-mode populations are then measured, yielding a set of time-averaged populations $\bar{\mathbf{n}}^{(i)}$. The weights $\mathbf{B}$ are then optimized over the training data so as to minimize the square error between the predictions of the photonic kernel machine $\hat{\mathbf{y}}$ and the known spectra of the noiseless pulses $\mathbf{y}$. Upon fine-tuning of the hyperparameter $\lambda$ by 10-fold cross validation, the optimal weights $\hat{\mathbf{B}}$ are obtained analytically as explained above. The error of the trained model is then evaluated on the test set. We here quantify this error by the relative absolute error on the energy $\Delta E/E_0$, where $\Delta E$ represents the energy area between the estimated and the actual energy-spectral-density curves of the original noiseless spectra. In

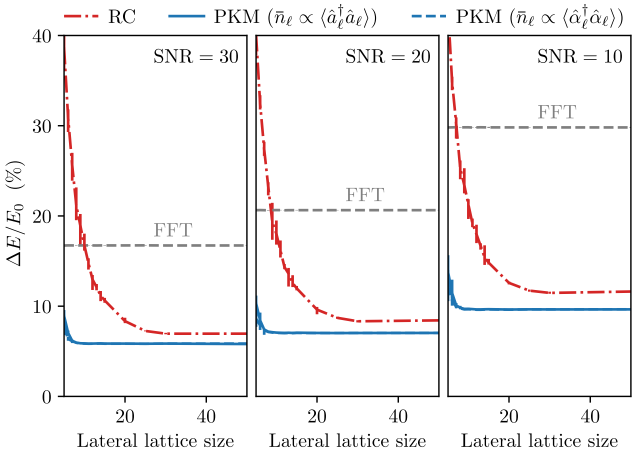



FIG. 6. Relative absolute error $\Delta E/E_0$ on the testing set for three values of the signal-to-noise ratio (SNR) for the fast Fourier transform (FFT, horizontal dashed line), the complex Ginzburg-Landau reservoir-computing (RC) model (dash-dotted), and the photonic kernel machine (PKM) where either the local-mode (plain) or the normal-mode (dashed) populations are measured. Data are averaged over five realizations of the reservoir; error bars correspond to the intervals between the lowest and highest errors over these realizations. For each realization, $N_{\text{train}} = 7000$ and $N_{\text{test}} = 3000$.

Fig. 6, we use this metric to benchmark our model against the ideal FFT of the noisy input signal and a the nonlinear polariton-based reservoir. For any of the chosen signal-to-noise ratio values and lattice sizes, the photonic kernel outperforms the other two approaches. Interestingly, the maximum performance is already reached with lattices as small as $10 \times 10$, with an error in the reproduced spectrum about three times smaller than that obtained by the ideal FFT procedure. In addition, Fig. 7 shows the error

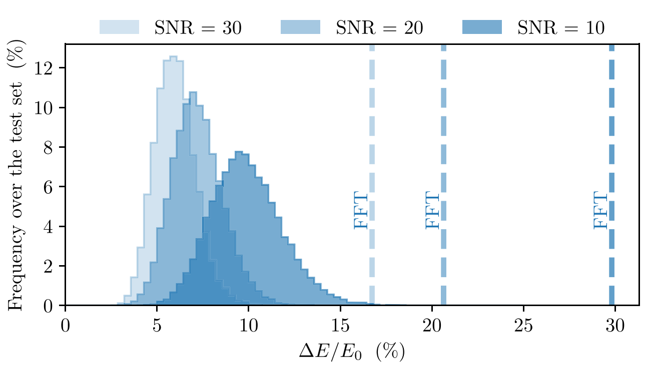



FIG. 7. Frequency of the error score over the testing set for three values of the signal-to-noise (SNR) ratio averaged over five realizations of a $20 \times 20$ photonic kernel machine. $N_{\text{test}} = 3000$ for each realization. The error score of the FFT is shown as dashed lines for comparison and falls systematically above the upper bound of that of the photonic kernel machine, in spite of being slower and requiring an ideal sampling rate significantly beyond the state of the art.

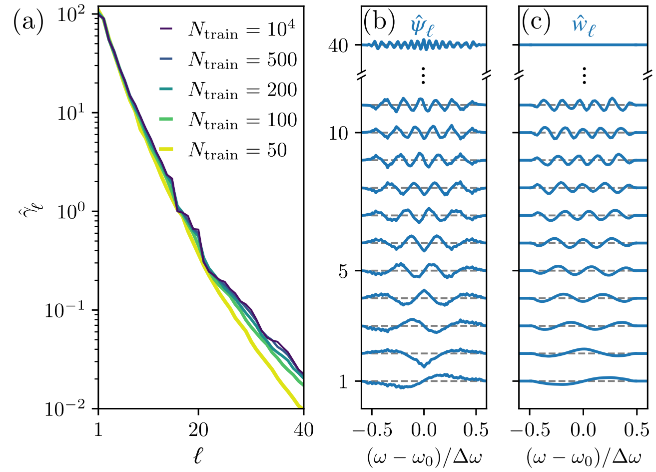



FIG. 8. (a) Convergence of the empirical spectrum of a photonic kernel for increasing training-set sizes. (b) First empirical eigenfunctions of the photonic kernel for $N_{\text{train}} = 10000$. (c) Functions learned by the photonic kernel machine during the optimization process. Parameters are $L = 20$, SNR $= 20$, and $\lambda = 10$.

frequency over the testing set for the three values of the signal-to-noise ratio and a $20 \times 20$ photonic lattice. Note that, for any amount of noise, the worst-case predictions remain more accurate than the ideal FFT.

In order to understand how the device learns from training examples, let us make use of the theoretical concepts introduced above. As expressed in Eq. (24), regression by the implemented photonic kernel machine is performed by a linear expansion $\hat{y}^{(i)}[\omega_n] = \sum_\ell \hat{w}_\ell[\omega_n]\langle\hat{\psi}_\ell|S^{(i)}\rangle$ of the feature-space components $\langle\hat{\psi}_\ell|S^{(i)}\rangle$ over a set of learned functions $\hat{w}_\ell$. The eigenvalues $\gamma_\ell$ and eigenfunctions $\psi_\ell/\sqrt{\gamma_\ell}$ of the photonic kernel do not depend on the optimization procedure and can be given empirical estimates from the measured populations $\bar{\mathbf{n}}^{(i)}$ during the training process, as described above. In Fig. 8(a), we show the convergence of the empirical eigenvalues as the number of training examples is increased, which is already reached for roughly $N_{\text{train}} = 1000$ with our training protocol. Figures 8(b) and 8(c) show the leading empirical feature maps as well as the corresponding learned eigenfunctions. In this task, one observes that the optimization procedure leads to a set of learned functions that correspond to filtered analogs of the empirical eigenfunctions of the kernel. One sees from Fig. 8(b) that the model builds a Fourier-sine expansion with a spectral resolution cutoff at $\sim \Delta\omega/2\pi L^2$. While in principle the optimization procedure may be sensitive to any feature-space component within this bound, the ridge regularization introduced during the optimization induces a soft cutoff for those whose associated eigenvalues have magnitudes lower than $\lambda$. This reduction to only the most statistically relevant components of the functional basis prevents the model from overfitting the training set, which would undermine its generalization capacity. In Fig. 8(c), the effect of the regularization is clearly visible on the highest-order represented learned function, which is completely filtered out. Hence, regularization here manifests itself as a low-pass filter on the learned decomposition.

From this figure, the interpretation of the photonic kernel machine regression mechanism becomes very clear: (i) the training set determines some optimal Fourier-like decomposition of the spectra; (ii) the regularization truncates the basis of such a decomposition to its most statistically relevant components, preventing the device from overfitting the noise; and, finally, (iii) the optimization finds a set of smooth filtered base functions on which to expand back those components to best reproduce the features. This results in the reconstruction of the noiseless spectrum, from which the noisy background uncorrelated with the training set is regularized out.

### B. Pulse shape recognition

We now show the performance of the same photonic kernel machine trained on a noisy-pulse shape classification problem.

The task involves determining whether the envelope of an input pulse is Gaussian ($y = 1$) or Lorentzian ($y = -1$) from the measurement of the photonic populations of the lattice in the presence of noise at the input port of the system. We prepare a set of pulses with central angular frequencies uniformly drawn at random in the band of interest $[\omega_0 - \Delta\omega/2, \omega_0 + \Delta\omega/2]$ and full widths at half maximum (FWHM) normally distributed around $\overline{\text{FWHM}} = \Delta\omega/20$ with a relative standard deviation of 10%. The populations induced are then time-integrated over some detection time $\Delta t$, yielding a vector of intensities $\bar{\mathbf{n}}$ that are finally linearly combined in such a way that the output of the photonic kernel machine is now a scalar of the form $\hat{f}(S) = \boldsymbol{\beta}^T\bar{\mathbf{n}}$. The optimization process is realized by minimizing the hinge loss $V(y^{(i)}, \hat{f}(S^{(i)})) = \max[0, 1 - y^{(i)}\hat{f}(S^{(i)})]$. This is achieved by means of a convex optimization solver [129] for $\lambda \to 0^+$.

The output of the $i$th pulse may equivalently be rewritten in terms of components of the feature map as $\hat{f}(S^{(i)}) = \hat{\mathbf{w}}^T\tilde{\mathbf{S}}^{(i)} + b$, with $\tilde{S}_\ell^{(i)} = \langle\psi_\ell|S^{(i)}\rangle$. It is worth noting that $\hat{f}(S) = \|\hat{\mathbf{w}}\| \times d(\tilde{\mathbf{S}})$, that is, the prediction for any incoming pulse $s$ is proportional to the signed distance $d(\tilde{\mathbf{S}})$ of its energy spectral density's feature-space coordinates to the plane $(\hat{\mathbf{w}}, b)$. As discussed above, this plane defines the feature-space decision boundary of the model. Input pulses are then classified into either of the two classes depending on whether their feature-space coordinates $\tilde{\mathbf{S}}^{(i)}$ fall on either side of this plane. Hence, predictions are of the form $\hat{y}^{(i)} = \text{sign}[\hat{f}(S^{(i)})]$.

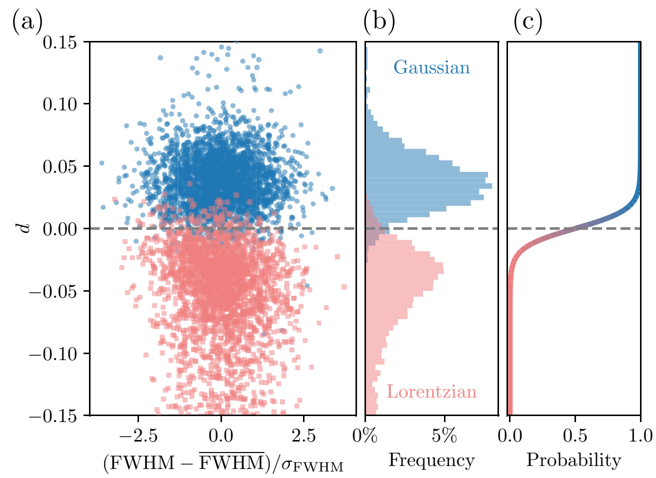
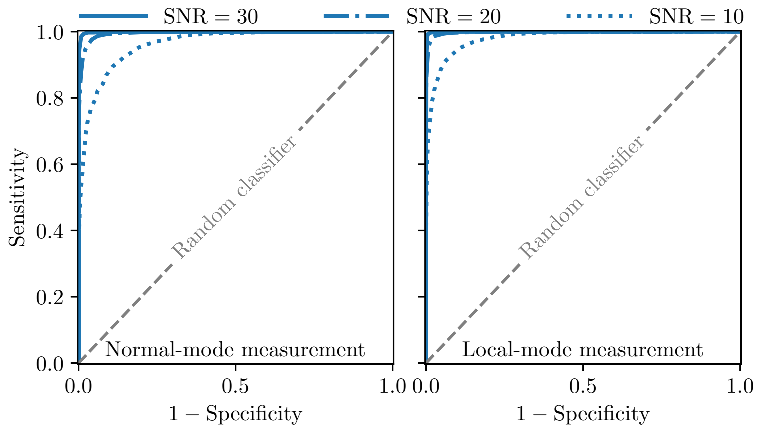



FIG. 9. (a) Signed distance from the features of the testing pulses to the learned hyperplane (decision boundary) as a function of their full widths at half maximum (FWHM) for a $20 \times 20$ photonic kernel machine under normal-mode measurement of the photonic populations. Dots correspond to Gaussian pulses, squares to Lorentzian ones. Pulses falling above the decision boundary (dashed line) are categorized as Gaussian by the classifier. Parameters: SNR = 20, $N_{\text{train}} = 14\,000$, and $N_{\text{test}} = 6000$. (b) Frequency of each class as a function of the signed distance to the discriminating hyperplane. (c) Probability that a pulse is Gaussian or Lorentzian at any given value of its associated signed distance to the discriminating hyperplane.

The classification process is illustrated in Fig. 9 for a $20 \times 20$ photonic kernel machine and intermediate noise strength (SNR = 20). In Fig. 9(a) one observes that features corresponding to either class indeed cluster at either of the sides of the discriminating hyperplane independently of the value of the FWHM. As becomes clear in Fig. 9(b), predictions become less accurate as the distance from the separating hyperplane becomes smaller, thereby giving an estimation of the likelihood of the prediction. This can be made more quantitative by calibrating the probability of the classifier. This probability is shown in Fig. 9(c) as given by $\mathbb{P}[s^{(i)} \text{ is Gaussian} \,|\, \hat{f}(S^{(i)})] = [1 + \exp(-A\hat{f}(S^{(i)}) + B)]^{-1}$, where the calibration parameters $A$ and $B$ were determined by Platt scaling [130].

The performance of a 20×20 trained photonic kernel machine is shown via its receiver operating characteristic (ROC) curve in Fig. 10 for increasing noise strengths and the two population measurement scenarios. This displays the sensitivity (true positive rate) as one allows the specificity of the model to drop (higher false positive rates) by playing on some external bias added to the trained model, the best tradeoff being found at the top left corner for no external bias. The ROC curve of an unbiased classifier that affects pulses randomly to either shape class is plotted as well for comparison. The high sensitivity and specificity




Normal-mode measurement

| SNR | sensitivity | specificity |
|---|---|---|
| 30 | 98.9% | 99.0% |
| 20 | 96.3% | 96.8% |
| 10 | 90.0% | 88.7% |

Local-mode measurement

| SNR | sensitivity | specificity |
|---|---|---|
| 30 | 99.5% | 99.3% |
| 20 | 97.9% | 98.3% |
| 10 | 92.7% | 91.9% |

FIG. 10. Receiver operating characteristic curve for a 20 × 20 photonic kernel machine on the pulse shape classification task for three noise strengths and two measurement scenarios. The closer to the top left corner, the better. The numerical values of their associated sensitivities and specificities are given in the tables on the right. Parameters: $N_{\text{train}} = 14000$ and $N_{\text{test}} = 6000$.

values of the trained classifier are given in the right panel of Fig. 10 for both population-measurement protocols.

### C. Frequency tracking

Only the case of pulsed input signals was investigated above. In what follows, we illustrate the performance of the photonic kernel machine presented above on the analysis of continuous rf signals by considering frequency estimation of noisy sinusoidal rf signals. This is of great relevance, for instance, in the context of short-time-scale force sensing with optomechanical devices [131,132]. In such applications, forces are read out by frequency tracking of the harmonic rf modulation imprinted on an optical signal by a resonating mechanical probe.

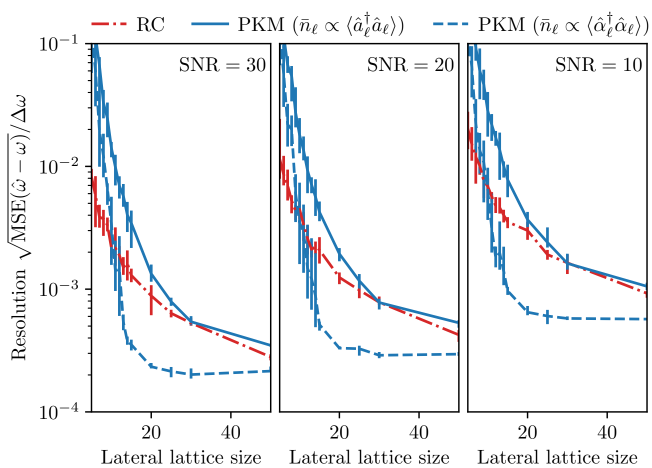



FIG. 11. Resolution of the photonic kernel machine (PKM) on the testing set for either local-mode (plain) or normal-mode (dashed) population measurements for increasing lattice sizes $L$ and three values of the signal-to-noise ratio (SNR). The performance of the nonlinear polariton-based reservoir-computing (RC) scheme (dash-dotted) is shown for comparison. Data is averaged over five realizations of the reservoir; error bars correspond to the intervals between the lowest and highest resolutions over these realizations. Parameters: $N_{\text{train}} = 7000$ and $N_{\text{test}} = 3000$.

To do so, we first generate a first set of training baseband signals consisting of complex exponentials with random initial phases and angular frequencies $\omega$ uniformly drawn between $\omega_0 - \Delta\omega/2$ and $\omega_0 + \Delta\omega/2$, to which white noise is added so as to match some signal-to-noise ratio SNR, here defined as the ratio between the average power of the sinusoidal baseband signal and that of the noise. We then measure the steady-state populations of the cavity resulting from the driving of the coupled cavities by the modulated signals, here after a time $\tau_d = 10/\bar{\kappa}$, and use this vector of populations to make frequency predictions of the form $\hat{f}(s) = \boldsymbol{\beta}^T \bar{\mathbf{n}}$. The vector $\boldsymbol{\beta}$ is then optimized so as to minimize the mean squared error between the estimated $\hat{\omega}^{(i)} = \hat{f}(s^{(i)})$ and the actual $\omega^{(i)}$ angular frequencies of all signals in the training set, with a ridge-regularization hyperparameter determined by 10-fold cross-validation.

The performance of the trained photonic kernel machine is finally evaluated on a testing set composed of new random complex exponentials. The average resolution achieved is shown in Fig. 11 as a function of the lattice size for increasing values of SNR, revealing the performance of the above-described photonic kernel machine on the spectral analysis of continuous rf signals. For cavities with $2\pi/\kappa \sim 10$ ps, the waiting time would be as little as $\tau_d \sim 10$ ps.

## VI. CONCLUSION

In this paper we presented photonic kernel machines, a framework for optical ultrafast spectral analysis of noisy rf signals that translates kernel methods from support-vector machines to photonic hardware. Such devices realize regression or classification tasks on high-dimensional data with above-gigahertz throughputs by utilizing the optical response of a set of optical modes to input analog signals as a measure of similarity.

We first gave a theoretical description of photonic kernel machines under very general assumptions. We analytically investigated the similarity kernel built into such devices and were able to express it explicitly from the susceptibility of the measured observables. Furthermore, we explored the feature maps associated to photonic kernel machines and found that their expressions could be experimentally determined from population measurements and the knowledge of the single-mode susceptibility.

We then studied a model describing a physical implementation consisting of a lattice of coupled linear optical cavities. We numerically demonstrated its capabilities on various regression and classification tasks, comprising the analysis of both pulsed and continuous rf signals. In particular, the proposed setup proved efficient in predicting the spectrum of picosecond pulses with nontrivial spectral structure from single-shot intensity measurements, being able to predict spectra with higher fidelity than the FFT of the noisy input signal. Moreover, it was shown to be able to discriminate pulses with distinct shapes as well as to estimate the angular frequency of continuous harmonic input signals. We showed that, by adding noise at the input of the device during the training protocol, the spurious effect of background noise on the predictive performance of the device could be successfully mitigated. On the spectrum estimation task, we could extract the actual feature maps associated to the simulated kernel machines as well as the basis of learned functions from which the photonic kernel machine composes its predictions. This allowed us to interpret the photonic kernel machine regression mechanism and revealed the ability of the system to filter out the uncorrelated background noise.

We believe that such devices, capable of analyzing above 1 million rf signals per second, may find applications in a broad variety of domains beyond spectroscopy. In the field of rf sensing, they could be used in pulse-Doppler radar systems as a way to optically analyze the reflected signals. In this way, the rate of emission could be increased by orders of magnitude, by relaxing the limiting dependence on the sampling rate of analog-to-digital converters. In telecommunications, these devices could be used, for instance, as a decoding means in frequency-shift-keying protocols involving high modulation rates in noisy environments. The fast frequency-tracking ability of the described device could be exploited in the context of short-time-scale force sensing with optomechanical resonators. Finally, photonic kernel machines could be integrated into more conventional machine-learning pipelines as a means of extracting nontrivial digital features from analog signals, acting as both an analog-to-digital converter and a preprocessing stage.

## ACKNOWLEDGMENTS

This work was supported by ERC Consolidator Grant NOMLI No. 770933, by ANR via the project UNIQ, and the FET flagship project PhoQuS (Grant Agreement ID No. 820392).

## APPENDIX: FEATURE-SPACE INSPECTION

The feature space embedding of kernel machines may be straightforwardly examined. Indeed, for a symmetric positive semidefinite kernel $K$, Mercer's theorem ensures that it admits an eigendecomposition of the form

$$K(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} \gamma_m \phi_m(\mathbf{x}) \phi_m(\mathbf{x}'), \tag{A1}$$

with $\gamma_{m+1} \leq \gamma_m$ and $\langle \phi_m, \phi_n \rangle_{p(\mathbf{x})} = \int d\mathbf{x} p(\mathbf{x}) \phi_m(\mathbf{x}) \phi_n(\mathbf{x}) = \delta_{m,n}$, where the measure is set to the probability density function of inputs $p$. The feature map of Eqs. (1) and (2) can then be easily shown to be given by $\psi_m(\mathbf{x}) = \sqrt{\gamma_m}\phi_m(\mathbf{x})$. We shall now identify the elements of this eigendecomposition.

To identify its eigenvalues and eigenvectors, one has to solve in principle for

$$\int d\mathbf{x}' p(\mathbf{x}') K(\mathbf{x}, \mathbf{x}') \phi(\mathbf{x}') = \gamma_m \phi_m(\mathbf{x}). \tag{A2}$$

Yet these may be estimated from the $N$ known training samples by replacing the input distribution $p(\mathbf{x})$ with the *empirical* distribution $\hat{p}(\mathbf{x})$, such that $\int d\mathbf{x} \hat{p}(\mathbf{x}) f(\mathbf{x}) = (1/N) \sum_i f(\mathbf{x}^{(i)})$ [114]. This yields a tractable discrete eigenvalue problem,

$$\frac{1}{N} \sum_{i=1}^{N} K(\mathbf{x}, \mathbf{x}^{(i)}) \hat{\phi}_m(\mathbf{x}^{(i)}) = \hat{\gamma}_m \hat{\phi}_m(\mathbf{x}), \tag{A3}$$

in terms of *empirical eigenvalues* $\hat{\gamma}_m$ and *empirical eigenfunctions* $\hat{\phi}_m$. It follows that the empirical eigenvalues correspond to the nonzero eigenvalues of either of the kernel matrices $\mathbf{K} = \mathbf{H}\mathbf{H}^T$ and $\mathbf{k} = \mathbf{H}^T\mathbf{H}$ [we recall that $H_{im} = h_m(\mathbf{x}^{(i)})$]. Indeed, because of their Gram matrix structure, their eigendecompositions,

$$\mathbf{K} = N\mathbf{U}\mathbf{D}_{\hat{\gamma}}\mathbf{U}^T, \quad \mathbf{k} = N\mathbf{u}\mathbf{d}_{\hat{\gamma}}\mathbf{u}^T, \tag{A4}$$

with $U_{im} = \hat{\phi}_m(\mathbf{x}^{(i)})/\sqrt{N}$, share the same nonzero eigenvalues and can be related through the following simple algebraic identity:

$$\mathbf{U}_m = \tfrac{1}{\sqrt{N\hat{\gamma}_m}} \mathbf{H}\mathbf{u}_m, \tag{A5}$$

following the singular-value decomposition of $\mathbf{K}$. In the physical model of the main text, one typically has $M \ll N$

and it becomes more suitable to work with the $M \times M$ matrix $\mathbf{k}$.

Similarly, the empirical eigenfunctions can be determined from Eq. (A3). By making use of the property of Eq. (A5), these finally read

$$\hat{\phi}_m(\mathbf{x}) = \frac{1}{\sqrt{N\hat{\gamma}_m}}\mathbf{u}_m^T\mathbf{h}(\mathbf{x}). \quad \text{(A6)}$$

From Eqs. (A1) and (A6), the kernel can be expanded into a series of empirical eigenfeature maps $K(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{M}\hat{\psi}_m(\mathbf{x})\hat{\psi}_m(\mathbf{x}')$, with orthogonal empirical feature maps simply given by

$$\hat{\psi}_m(\mathbf{x}) = \sqrt{\hat{\gamma}_m}\hat{\phi}_m(\mathbf{x}). \quad \text{(A7)}$$

Thus, the internal functional representations of the model are *independent* of the optimization process, completely determined by the statistics of the training samples, and can be estimated by diagonalizing the matrix $\mathbf{k}$.

The predictions of the model in the abstract feature-space picture, $\hat{f}(\mathbf{x}) = \mathbf{w}^T\tilde{\mathbf{x}} \equiv \mathbf{w}^T\boldsymbol{\psi}(\mathbf{x})$, take the form of a linear combination of features whose weights, determined through training, can be geometrically interpreted as the parameters of a hyperplane. As a function of the primal parameters, this learned hyperplane is characterized by

$$\hat{\mathbf{w}} = \sqrt{N}\mathbf{u}^T\hat{\boldsymbol{\beta}}. \quad \text{(A8)}$$

Therefore, given a set of generating functions $\{h_m\}_m$, one is now able to access to a complete understanding of the model and its feature-space representations.


[1] P. Ambs, Optical computing: A 60-year adventure, Adv. Opt. Technol. **2010**, e372652 (2010).
[2] P. Mengert and T. T. Tanimoto, Control apparatus (1966), US Patent 35, 258, 56.
[3] C. S. Weaver and J. W. Goodman, A technique for optically convolving two functions, Appl. Opt. **5**, 1248 (1966).
[4] J. W. Goodman, A. R. Dias, and L. M. Woody, Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms, Opt. Lett. **2**, 1 (1978).
[5] A. Sawchuk and T. Strand, Digital optical computing, Proc. IEEE **72**, 758 (1984).
[6] A. Huang, Architectural considerations involved in the design of an optical digital computer, Proc. IEEE **72**, 780 (1984).
[7] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, Artif. Intell Rev. **53**, 5455 (2020).
[8] M. M. Waldrop, The chips are down for moore's law, Nat. News **530**, 144 (2016).
[9] E. Strubell, A. Ganesh, and A. McCallum, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, 2019), p. 3645.
[10] J. Dong, S. Gigan, F. Krzakala, and G. Wainrib, in *2018 IEEE Statistical Signal Processing Workshop (SSP)* (2018), p. 448.
[11] R. Ohana, J. Wacker, J. Dong, S. Marmin, F. Krzakala, M. Filippone, and L. Daudet, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), p. 9294.
[12] D. Schneider, Lightmatter's Mars Chip Performs Neural-Network Calculations at the Speed of Light: MIT spinoff harnesses optical computing to make neural-networks run faster and more efficiently, IEEE Spectrum (2020).
[13] D. Schneider, Deep learning at the speed of light: Lightmatter bets that optical computing can solve AI's efficiency problem, IEEE Spectr. **58**, 28 (2021).
[14] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis, Inference in artificial intelligence with deep optics and photonics, Nature **588**, 39 (2020).
[15] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, All-optical machine learning using diffractive deep neural networks, Science **361**, 1004 (2018).
[16] M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, and V. J. Sorger, Massively parallel amplitude-only Fourier neural network, Optica **7**, 1812 (2020).
[17] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, Parallel convolutional processing using an integrated photonic tensor core, Nature **589**, 52 (2021).
[18] M. D. Zeiler and R. Fergus, in *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Springer International Publishing, Cham, 2014), p. 818.
[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).
[20] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, Extreme learning machine: Theory and applications, Neurocomput. Neural Networks **70**, 489 (2006).
[21] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst., Man, and Cybernetics, Part B (Cybernetics) **42**, 513 (2012).
[22] H. Jaeger, The "echo state" approach to analysing and training recurrent neural networks, Bonn, Germany: German Nat. Res. Center Inf. Technol. GMD Tech. Rep. **148**, 34 (2001).
[23] H. Jaeger and H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science **304**, 78 (2004).
[24] W. Maass, T. Natschläger, and H. Markram, Real-time computing without stable states: A new framework for neural computation based on perturbations, Neural Comput. **14**, 2531 (2002).
[25] D. Verstraeten, B. Schrauwen, D. Stroobandt, and J. Van Campenhout, Isolated word recognition with the liquid state machine: A case study, Inf. Process. Lett. Appl. Spiking Neural Networks **95**, 521 (2005).

[26] D. Verstraeten, B. Schrauwen, and D. Stroobandt, Reservoir-based techniques for speech recognition, The 2006 IEEE International Joint Conference on Neural Network Proceedings (2006).

[27] M. Lukoševičius and H. Jaeger, Reservoir computing approaches to recurrent neural network training, Comput. Sci. Rev. **3**, 127 (2009).

[28] A. Rodan and P. Tino, Minimum complexity echo state network, IEEE Trans. Neural Networks **22**, 131 (2011).

[29] 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 Netw. **115**, 100 (2019).

[30] K. Nakajima, Physical reservoir computing–an introductory perspective, Jpn. J. Appl. Phys. **59**, 060501 (2020).

[31] G. Marcucci, D. Pierangeli, and C. Conti, Theory of Neuromorphic Computing by Waves: Machine Learning by Rogue Waves, Dispersive Shocks, and Solitons, Phys. Rev. Lett. **125**, 093901 (2020).

[32] G. V. der Sande, D. Brunner, and M. C. Soriano, Advances in photonic reservoir computing, Nanophotonics **6**, 561 (2017).

[33] S. Sunada, K. Kanno, and A. Uchida, Using multidimensional speckle dynamics for high-speed, large-scale, parallel photonic computing, Opt. Express **28**, 30349 (2020).

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

[35] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, Experimental demonstration of reservoir computing on a silicon photonics chip, Nat. Commun. **5**, 3541 (2014).

[36] F. Denis-Le Coarer, M. Sciamanna, A. Katumba, M. Freiberger, J. Dambre, P. Bienstman, and D. Rontani, All-optical reservoir computing on a photonic chip using silicon-based ring resonators, IEEE J. Sel. Top. Quantum Electron. **24**, 1 (2018).

[37] M. S. Kulkarni and C. Teuscher, in *2012 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)* (2012), p. 226.

[38] C. Du, F. Cai, M. A. Zidan, W. Ma, S. H. Lee, and W. D. Lu, Reservoir computing using dynamic memristors for temporal information processing, Nat. Commun. **8**, 2204 (2017).

[39] S. Boyn, J. Grollier, G. Lecerf, B. Xu, N. Locatelli, S. Fusil, S. Girod, C. Carrétéro, K. Garcia, S. Xavier, J. Tomas, L. Bellaiche, M. Bibes, A. Barthélémy, S. Saïghi, and V. Garcia, Learning through ferroelectric domain dynamics in solid-state synapses, Nat. Commun. **8**, 14736 (2017).

[40] R. Nakane, G. Tanaka, and A. Hirose, Reservoir computing with spin waves excited in a garnet film, IEEE Access **6**, 4462 (2018).

[41] D. Marković, N. Leroux, M. Riou, F. Abreu Araujo, J. Torrejon, D. Querlioz, A. Fukushima, S. Yuasa, J. Trastoy, P. Bortolotti, and J. Grollier, Reservoir computing with the frequency, phase, and amplitude of spin-torque nano-oscillators, Appl. Phys. Lett. **114**, 012409 (2019).

[42] K. Vandoorne, W. Dierckx, B. Schrauwen, D. Verstraeten, R. Baets, P. Bienstman, and J. V. Campenhout, Toward optical signal processing using photonic reservoir computing, Opt. Express **16**, 11182 (2008).

[43] Y. Paquot, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, in *Nonlinear Optics and Applications IV*, Vol. 7728 (International Society for Optics and Photonics, 2010), p. 77280B.

[44] L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, Information processing using a single dynamical node as complex system, Nat. Commun. **2**, 468 (2011).

[45] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman, Parallel reservoir computing using optical amplifiers, IEEE Trans. Neural Networks **22**, 1469 (2011).

[46] F. Duport, B. Schneider, A. Smerieri, M. Haelterman, and S. Massar, All-optical reservoir computing, Opt. Express **20**, 22783 (2012).

[47] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, Optoelectronic reservoir computing, Sci. Rep. **2**, 287 (2012).

[48] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, Photonic information processing beyond turing: An optoelectronic implementation of reservoir computing, Opt. Express **20**, 3241 (2012).

[49] M. C. Soriano, S. Ortín, D. Brunner, L. Larger, C. R. Mirasso, I. Fischer, and L. Pesquera, Optoelectronic reservoir computing: Tackling noise-induced performance degradation, Opt. Express **21**, 12 (2013).

[50] R. M. Nguimdo, G. Verschaffelt, J. Danckaert, and G. Van der Sande, Simultaneous computation of two independent tasks using reservoir computing based on a single photonic nonlinear node with optical feedback, IEEE Trans. Neural Netw. Learn. Syst. **26**, 3301 (2015).

[51] S. Ortín, M. C. Soriano, L. Pesquera, D. Brunner, D. San-Martín, I. Fischer, C. R. Mirasso, and J. M. Gutiérrez, A unified framework for reservoir computing and extreme learning machines based on a single time-delayed neuron, Sci. Rep. **5**, 14945 (2015).

[52] L. Larger, A. Baylón-Fuentes, R. Martinenghi, V. S. Udaltsov, Y. K. Chembo, and M. Jacquot, High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification, Phys. Rev. X **7**, 011015 (2017).

[53] Y. K. Chembo, Machine learning based on reservoir computing with time-delayed optoelectronic and photonic systems, Chaos **30**, 013111 (2020).

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

[55] 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).

[56] R. Mirek, A. Opala, P. Comaron, M. Furman, M. Król, K. Tyszka, B. Seredyński, D. Ballarini, D. Sanvitto,

T. C. H. Liew, W. Pacuski, J. Suffczyński, J. Szczytko, M. Matuszewski, and B. Piętka, Neuromorphic binarized polariton networks, Nano Lett. **21**, 3715 (2021).

[57] J. Dong, M. Rafayelyan, F. Krzakala, and S. Gigan, Optical reservoir computing using multiple light scattering for chaotic systems prediction, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).

[58] M. Rafayelyan, J. Dong, Y. Tan, F. Krzakala, and S. Gigan, Large-Scale Optical Reservoir Computing for Spatiotemporal Chaotic Systems Prediction, Phys. Rev. X **10**, 041037 (2020).

[59] A. Sengupta, Y. Shim, and K. Roy, Proposal for an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets, IEEE Trans. Biomed. Circuits Syst. **10**, 1152 (2016).

[60] J. Grollier, D. Querlioz, and M. D. Stiles, Spintronic nanodevices for bioinspired computing, Proc. IEEE **104**, 2024 (2016).

[61] J. Torrejon, M. Riou, F. A. Araujo, S. Tsunegi, G. Khalsa, D. Querlioz, P. Bortolotti, V. Cros, K. Yakushiji, A. Fukushima, H. Kubota, S. Yuasa, M. D. Stiles, and J. Grollier, Neuromorphic computing with nanoscale spintronic oscillators, Nature **547**, 428 (2017).

[62] M. Romera, P. Talatchian, S. Tsunegi, F. Abreu Araujo, V. Cros, P. Bortolotti, J. Trastoy, K. Yakushiji, A. Fukushima, H. Kubota, S. Yuasa, M. Ernoult, D. Vodenicarevic, T. Hirtzlin, N. Locatelli, D. Querlioz, and J. Grollier, Vowel recognition with four coupled spin-torque nanooscillators, Nature **563**, 230 (2018).

[63] J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, Neuromorphic spintronics, Nat. Electron. **3**, 360 (2020).

[64] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, Nanoscale memristor device as synapse in neuromorphic systems, Nano Lett. **10**, 1297 (2010).

[65] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation, IEEE Trans. Electron. Devices **58**, 2729 (2011).

[66] Y. Yang, B. Chen, and W. D. Lu, Memristive physically evolving networks enabling the emulation of heterosynaptic plasticity, Adv. Mater. **27**, 7720 (2015).

[67] S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, and W. D. Lu, Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity, Nano Lett. **15**, 2203 (2015).

[68] M. Chu, B. Kim, S. Park, H. Hwang, M. Jeon, B. H. Lee, and B.-G. Lee, Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron, IEEE Trans. Ind. Electron. **62**, 2410 (2015).

[69] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, Training and operation of an integrated neuromorphic network based on metal-oxide memristors, Nature **521**, 61 (2015).

[70] D. S. Jeong, K. M. Kim, S. Kim, B. J. Choi, and C. S. Hwang, Memristors for energy-efficient new computing paradigms, Adv. Electron. Mater. **2**, 1600090 (2016).

[71] C.-C. Hsieh, A. Roy, Y.-F. Chang, D. Shahrjerdi, and S. K. Banerjee, A sub-1-volt analog metal oxide memristive-based synaptic device with large conductance change for energy-efficient spike-based computing systems, Appl. Phys. Lett. **109**, 223501 (2016).

[72] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Wu, M. Barnell, G.-L. Li, H. L. Xin, R. S. Williams, Q. Xia, and J. J. Yang, Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing, Nat. Mater. **16**, 101 (2017).

[73] J. H. Yoon, Z. Wang, K. M. Kim, H. Wu, V. Ravichandran, Q. Xia, C. S. Hwang, and J. J. Yang, An artificial nociceptor based on a diffusive memristor, Nat. Commun. **9**, 417 (2018).

[74] Z. Wang, *et al.*, Fully memristive neural networks for pattern classification with unsupervised learning, Nat. Electron. **1**, 137 (2018).

[75] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, Efficient and self-adaptive in-situ learning in multilayer memristor neural networks, Nat. Commun. **9**, 2385 (2018).

[76] Y. Kim, Y. J. Kwon, D. E. Kwon, K. J. Yoon, J. H. Yoon, S. Yoo, H. J. Kim, T. H. Park, J.-W. Han, K. M. Kim, and C. S. Hwang, Nociceptive memristor, Adv. Mater. **30**, 1704320 (2018).

[77] D. S. Jeong and C. S. Hwang, Nonvolatile memory materials for neuromorphic intelligent machines, Adv. Mater. **30**, 1704729 (2018).

[78] A. Afifi, A. Ayatollahi, and F. Raissi, in *2009 European Conference on Circuit Theory and Design* (2009), p. 563.

[79] D. Querlioz, O. Bichler, and C. Gamrat, in *The 2011 International Joint Conference on Neural Networks* (2011), p. 1775.

[80] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, Immunity to device variations in a spiking neural network with memristive nanodevices, IEEE Trans. Nanotechnol. **12**, 288 (2013).

[81] G. Srinivasan, A. Sengupta, and K. Roy, Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip STDP learning, Sci. Rep. **6**, 29545 (2016).

[82] K. Roy, A. Jaiswal, and P. Panda, Towards spike-based machine intelligence with neuromorphic computing, Nature **575**, 607 (2019).

[83] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, Going deeper in spiking neural networks: VGG and residual architectures, Front. Neurosci. **13**, 95 (2019).

[84] M. Ernoult, J. Grollier, D. Querlioz, Y. Bengio, and B. Scellier, Updates of Equilibrium Prop Match Gradients of Backprop Through Time in an RNN with Static Input, (2019), ArXiv:1905.13633.

[85] M. Ernoult, J. Grollier, D. Querlioz, Y. Bengio, and B. Scellier, Equilibrium Propagation with Continual Weight Updates, (2020), ArXiv:2005.04168.

[86] A. Laborieux, M. Ernoult, B. Scellier, Y. Bengio, J. Grollier, and D. Querlioz, Scaling equilibrium propagation to deep convNets by drastically reducing its gradient estimator bias, Front. Neurosci. **15**, 633674 (2021).

[87] J. Laydevant, M. Ernoult, D. Querlioz, and J. Grollier, Training Dynamical Binary Neural Networks with Equilibrium Propagation, (2021), ArXiv:2103.08953.
[88] E. Martin, M. Ernoult, J. Laydevant, S. Li, D. Querlioz, T. Petrisor, and J. Grollier, EqSpike: Spike-driven equilibrium propagation for neuromorphic implementations, iScience **24**, 102222 (2021).
[89] In the recent years, several works have also addressed this matter from a privacy perspective, proposing quantum-secure machine-learning protocols [133–136].
[90] S. Ghosh, A. Opala, M. Matuszewski, T. Paterek, and T. C. H. Liew, Quantum reservoir processing, npj Quantum Inf. **5**, 1 (2019).
[91] S. Ghosh, A. Opala, M. Matuszewski, T. Paterek, and T. C. H. Liew, Reconstructing Quantum States With Quantum Reservoir Networks, IEEE Trans. Neural Networks Learn. Syst., 1 (2020).
[92] S. Ghosh, T. Paterek, and T. C. H. Liew, Quantum Neuromorphic Platform for Quantum State Preparation, Phys. Rev. Lett. **123**, 260404 (2019).
[93] S. Ghosh, T. Krisnanda, T. Paterek, and T. C. H. Liew, Realising and compressing quantum circuits with quantum reservoir computing, Commun. Phys. **4**, 1 (2021).
[94] T. Krisnanda, S. Ghosh, T. Paterek, and T. C. H. Liew, Creating and concentrating quantum resource states in noisy environments using a quantum neural network, Neural Netw. **136**, 141 (2021).
[95] 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).
[96] D. Marković and J. Grollier, Quantum neuromorphic computing, Appl. Phys. Lett. **117**, 150501 (2020).
[97] B. E. Boser, I. M. Guyon, and V. N. Vapnik, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92 (Association for Computing Machinery, New York, NY, USA, 1992), p. 144.
[98] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, Massachusetts, USA, 2001).
[99] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Springer Series in Statistics* (Springer New York, New York, 2009).
[100] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, USA, 2016).
[101] Note that this in general is not a diffeomorphism and thus *stricto sensu* not a change of coordinates.
[102] A. Caponnetto and E. De Vito, Optimal rates for the regularized least-squares algorithm, Found. Comput. Math. **7**, 331 (2007).
[103] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, edited by F. Bach, Adaptive Computation and Machine Learning Series (MIT Press, Cambridge, MA, USA, 2012).
[104] M. J. D. Powell, Radial basis functions for multivariable interpolation : A review, Algorithms for Approximation (1987).
[105] D. S. Broomhead and D. Lowe, Multivariable functional interpolation and adaptive networks, Complex Syst. **2**, 321 (1988).
[106] T. van Gestel, J. A. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. de Moor, and J. Vandewalle, Benchmarking least squares support vector machine classifiers, Mach. Learn. **54**, 5 (2004).
[107] N. Liu and P. Rebentrost, Quantum machine learning for quantum anomaly detection, Phys. Rev. A **97**, 042315 (2018).
[108] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, Phys. Rev. Lett. **122**, 040504 (2019).
[109] 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).
[110] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature **567**, 209 (2019).
[111] 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**, 1 (2021).
[112] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, (2020), ArXiv:2010.02174.
[113] M. Schuld, Supervised quantum machine learning models are kernel methods, (2021), ArXiv:2101.11020.
[114] C. Williams and M. Seeger, in *Advances in Neural Information Processing Systems 13* (MIT Press, 2001), p. 682.
[115] P. Drineas and M. W. Mahoney, On the nystrom method for approximating a gram matrix for improved kernel-based learning, J. Mach. Learn. Res. **6**, 2153 (2005).
[116] A. Rahimi and B. Recht, in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07 (Curran Associates Inc., Red Hook, NY, USA, 2007), p. 1177.
[117] A. Rahimi and B. Recht, Weighted sums of random kitchen sinks: in *Advances in Neural Information Processing Systems*, Vol. 21, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Curran Associates, Inc., 2009).
[118] A. Saade, F. Caltagirone, I. Carron, L. Daudet, A. Dremeau, S. Gigan, and F. Krzakala, Random projections through multiple optical scattering: Approximating Kernels at the speed of light, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2016).
[119] For simplicity, no bias is here considered, without loss of generality as one may always write $\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} \beta_m h_m(\mathbf{x}) + b = \sum_{m=0}^{M} \beta_m h_m(\mathbf{x})$, with $h_0 = 1$ and $\beta_0 = b$.
[120] B. Schölkopf, A. Smola, and K.-R. Müller, in *Advances in Kernel Methods - Support Vector Learning* (MIT Press, 1999), p. 327.
[121] S. Mika, B. Schölkopf, A. Smola, K. R. Müller, M. Scholz, and G. Rätsch, in *Advances in Neural Information Processing Systems 11 - Proceedings of the 1998 Conference, NIPS 1998* (Neural information processing systems foundation, 1999), p. 536.
[122] See for instance Texas Instruments' ADC12DL3200 (6.4 GSPS) or Analog Devices' AD9213 (10.25 GSPS).

[123] C. W. Gardiner and M. J. Collett, Input and output in damped quantum systems: Quantum stochastic differential equations and the master equation, Phys. Rev. A **31**, 3761 (1985).
[124] C. Gardiner and P. Zoller, *Quantum Noise: A Handbook of Markovian and Non-Markovian Quantum Stochastic Methods with Applications to Quantum Optics*, third edition ed., Springer Series in Synergetics (Springer-Verlag, Berlin, Heidelberg, 2004).
[125] $K$ and $\mathcal{K}$ may be thought of as components of the same operator $\hat{K} = \sum_\ell |h_\ell\rangle\langle h_\ell|$. Indeed, one has $K(S, S') = \langle S|\hat{K}|S'\rangle$ and $\mathcal{K}(\omega, \omega') = \langle\omega|\hat{K}|\omega'\rangle$.
[126] R. Zhang and W. Wang, Facilitating the applications of support vector machine by using a new kernel, Expert Syst. Appl. **38**, 14225 (2011).
[127] In practice, features $\bar{\mathbf{n}}$ are first scaled using the z-score standardization prescription: $\bar{n}_\ell \mapsto (\bar{n}_\ell - \mathbb{E}_{\hat{p}}[\bar{n}_\ell])/\sqrt{\mathrm{Var}_{\hat{p}}[\bar{n}_\ell]}$.
[128] S. Rosset, J. Zhu, and T. Hastie, in *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03 (MIT Press, Cambridge, MA, USA, 2003), p. 1237.
[129] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd, in *Proceedings of the 1st First Workshop for High Performance Technical Computing in Dynamic Languages*, HPTCDL '14 (IEEE Press, New Orleans, Louisiana, 2014), p. 18.
[130] J. C. Platt, in *Advances in Large Margin Classifiers* (MIT Press, 1999), p. 61.
[131] P. E. Allain, L. Schwab, C. Mismer, M. Gely, E. Mairiaux, M. Hermouet, B. Walter, G. Leo, S. Hentz, M. Faucher, G. Jourdan, B. Legrand, and I. Favero, Optomechanical resonating probe for very high frequency sensing of atomic forces, Nanoscale **12**, 2939 (2020).
[132] B. Guha, P. E. Allain, A. Lemaître, G. Leo, and I. Favero, Force Sensing with an Optomechanical Self-Oscillator, Phys. Rev. Appl. **14**, 024079 (2020).
[133] J. Bang, S.-W. Lee, and H. Jeong, Protocol for secure quantum machine learning at a distant place, Quantum Inf. Process. **14**, 3933 (2015).
[134] Y.-B. Sheng and L. Zhou, Distributed secure quantum machine learning, Sci. Bull. **62**, 1025 (2017).
[135] W. Song, Y. Lim, H. Kwon, G. Adesso, M. Wieśniak, M. Pawłowski, J. Kim, and J. Bang, Quantum secure learning with classical samples, Phys. Rev. A **103**, 042409 (2021).
[136] W. Li, S. Lu, and D.-L. Deng, Quantum federated learning through blind quantum computing, Sci. China Phys., Mech. Astron. **64**, 100312 (2021).