

PAPER • OPEN ACCESS

# Defence against adversarial attacks using classical and quantum-enhanced Boltzmann machines<sup>†</sup>

To cite this article: Aidan Kehoe *et al* 2021 *Mach. Learn.: Sci. Technol.* **2** 045006

View the [article online](#) for updates and enhancements.

## You may also like

- [Extracting electron scattering cross sections from swarm data using deep neural networks](#)  
Vishrut Jetly and Bhaskar Chaudhury
- [Data-driven discovery of Koopman eigenfunctions for control](#)  
Eurika Kaiser, J Nathan Kutz and Steven L Brunton
- [Adaptive partial scanning transmission electron microscopy with reinforcement learning](#)  
Jeffrey M Ede



## PAPER

## OPEN ACCESS



RECEIVED  
21 December 2020REVISED  
16 March 2021ACCEPTED FOR PUBLICATION  
30 March 2021PUBLISHED  
15 July 2021

Original Content from  
this work may be used  
under the terms of the  
[Creative Commons  
Attribution 4.0 licence](#).

Any further distribution  
of this work must  
maintain attribution to  
the author(s) and the title  
of the work, journal  
citation and DOI.



# Defence against adversarial attacks using classical and quantum-enhanced Boltzmann machines<sup>†</sup>

Aidan Kehoe<sup>1,\*</sup>, Peter Wittek<sup>2,3,4,5</sup>, Yanbo Xue<sup>6,\*</sup>  and Alejandro Pozas-Kerstjens<sup>7,\*</sup> <sup>1</sup> GraphStax Research, Toronto, Canada<sup>2</sup> Rotman School of Management—University of Toronto, M5S 3E6 Toronto, Canada<sup>3</sup> Creative Destruction Lab, M5S 3E6 Toronto, Canada<sup>4</sup> Vector Institute for Artificial Intelligence, M5G 1M1 Toronto, Canada<sup>5</sup> Perimeter Institute for Theoretical Physics, N2L 2Y5 Waterloo, Canada<sup>6</sup> Career Science Lab—BOSS Zhipin, Beijing, China<sup>7</sup> Department of Mathematical Analysis—Universidad Complutense de Madrid, 28040 Madrid, Spain

\* Authors to whom any correspondence should be addressed.

E-mail: [akehoe@graphstax.com](mailto:akehoe@graphstax.com), [xueyanbo@kanzhun.com](mailto:xueyanbo@kanzhun.com) and [physics@alexpozas.com](mailto:physics@alexpozas.com)**Keywords:** generative models, Boltzmann machines, quantum machine learning, adversarial attacks, machine learning security

## Abstract

We provide a robust defence to adversarial attacks on discriminative algorithms. Neural networks are naturally vulnerable to small, tailored perturbations in the input data that lead to wrong predictions. On the contrary, generative models attempt to learn the distribution underlying a dataset, making them inherently more robust to small perturbations. We use Boltzmann machines for discrimination purposes as attack-resistant classifiers, and compare them against standard state-of-the-art adversarial defences. We find improvements ranging from 5% to 72% against attacks with Boltzmann machines on the MNIST dataset. We furthermore complement the training with quantum-enhanced sampling from the D-Wave 2000Q annealer, finding results comparable with classical techniques and with marginal improvements in some cases. These results underline the relevance of probabilistic methods in constructing neural networks and highlight a novel scenario of practical relevance where quantum computers, even with limited hardware capabilities, could provide advantages over classical computers.

## 1. Introduction

The robustness of machine learning models against adversarial attacks is currently an open question. Indeed, it is easy to make neural networks misclassify images by applying perturbations to just one single pixel [1], or even perturbations that are imperceptible to human observers. Such a perturbation is found by maximizing the network's prediction error [2, 3], therefore showing adversarial examples which rely on the inherent uncertainty that neural networks have about their predictions [4]. These ideas lead to real-world attempts at fooling state-of-the-art computer vision models [5] and enabling malware to bypass detection [6, 7].

Right now there exist a great plethora of adversarial attacks on neural-network discriminative algorithms. The most common are white-box, meaning that they have access to the neural network parameters after training [2, 3, 8, 9]. In contrast, black-box attacks just query the classifier to construct adversarial examples [10, 11]. These latter attacks are more difficult to create but potentially more efficient to a wide range of classifiers, not just deep neural networks. However, it has been shown that substitute models can be trained on black-box queries to a target model, and adversarial examples built from white-box attacks to the substitute model are also effective in the target [12]. It is hotly debated whether there is a type of machine learning model, or any data processing technique, that is most adept at defending against adversarial examples. Indeed, right now defences are often broken soon after their publication [13–16], and part of the

<sup>†</sup> This work is dedicated to the memory of Peter Wittek.

community's proposal is encouraging a faster iteration cycle between attacks and defences via competitions [17].

Bayesian methods and generative probabilistic graphical models (PGMs) could potentially be more robust, since their training is not necessarily related to backpropagation and first-order gradient descent. Gaussian processes, for instance, achieved a remarkable performance against attacks [18, 19]. In a similar way, generative models, which learn the full joint probability distribution of instances and labels instead of an input–output function, are thought to be more resistant to attacks based on small perturbations. The fact that the full data manifold is approximated in generative models was the inspiration behind generative adversarial networks (GANs) [20], that were originally designed such that a generator would approximate the data manifold to fool a discriminator. Our motivation stems from this observation and we ask the question whether generative probabilistic models can have an improved robustness to adversarial attacks when compared against feedforward neural networks.

In this work we benchmark the robustness of generative PGMs against recent state-of-the-art attacks on discriminative neural networks, and assess the performance of state-of-the-art defences on these attacks. We compare the results previously known for feedforward deep classifiers against classical and quantum-enhanced restricted Boltzmann machines (RBMs) in white-box attack schemes. Our findings show that both classical and quantum-enhanced Boltzmann machines far outperform the current competition, with improvements ranging from 4.62% to 72.41%. Moreover, quantum-enhanced Boltzmann machines give marginal improvements over their classical counterparts in some instances, which is an encouraging result given the young state of quantum hardware platforms, and illustrates a practical application of interest of noisy, intermediate-scale quantum technologies more direct than demonstrations of quantum superiority [21, 22].

This paper is organized as follows: in section 2 we introduce the concept of adversarial attacks and review the relevant recent work in this area. We proceed analogously in section 3 for the training of RBMs. Then, in section 4 we describe the particular architectures, attacks and defences we consider, and evaluate them in section 5. We conclude in section 6 with a discussion and pointing to directions of future work.

## 2. Adversarial examples

Formally, an adversarial attack is aimed at a classifier  $f: \mathcal{X} \mapsto \mathcal{Y}$ , where  $\mathcal{X}$  is typically a subspace of  $\mathbb{R}^d$ , and  $\mathcal{Y}$  is a discrete space corresponding to labels. The goal of the attack is to create an example  $\mathbf{x}^*$  in the vicinity of a valid training point  $\mathbf{x}$  such that  $\|\mathbf{x}^* - \mathbf{x}\|$  is small and  $f(\mathbf{x}^*) \neq f(\mathbf{x})$ . If  $f(\mathbf{x}^*)$  does not have any restriction other than  $f(\mathbf{x}^*) \neq f(\mathbf{x})$ , the attack is said to be non-targeted. In contrast, in a targeted attack the adversary aims to specify the label of the adversarial example, that is,  $f(\mathbf{x}^*) = y^*$  for a specific choice of  $y^*$ . The difference to the actual examples should be small as measured by the  $L_p$  norm, where  $p$  is typically 0, 1, 2, or  $\infty$ . For a more complete taxonomy, see [10, 17, 23].

As an illustration of a standard attack, let us consider the fast gradient sign method (FGSM) [3]. This attack is interesting since it approximates the minimization in the infinity norm bound,  $\|\mathbf{x}^* - \mathbf{x}\|_\infty$ , in a single step, which means that it can scale easily to large datasets. FGSM is an example of a white-box attack, where the model parameters are known to the adversary. Adversarial examples are created by

$$\mathbf{x}^* = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}_\theta(\mathbf{x}, y)), \quad (1)$$

where  $\nabla_{\mathbf{x}} \mathcal{L}_\theta(\mathbf{x}, y)$  is the gradient of the loss function  $\mathcal{L}_\theta$  of the machine learning model parametrized by  $\theta$  in  $\mathbf{x}$ , whose true label is  $y$ . In its simplest formulation, FGSM does not transfer well to other models than the one it was trained on because of the fact that all the information needed to prepare the adversarial examples is very specific to the model used. On the other hand, black-box schemes in turn either transfer attacks from an undefended model to other models or query an unknown model to gain information about its decision boundary. In both cases, the agnosticity about the target model make black-box attacks more difficult to craft but potentially more 'dangerous' in terms of their reach.

Adversarial examples in linear models can be explained as a result of high-dimensional dot products. In such models, the output when a perturbed image is input will be the sum of the dot product of the weights with the unperturbed input and the dot product of the weights with the perturbation. If each weight vector has average magnitude  $m$  and the input is  $n$ -dimensional, the activation in the classifier will grow by  $\epsilon mn$  (where  $\epsilon$  is the magnitude of the perturbation). For low-dimensional problems this does not result in a large difference in classification between the original and perturbed data, as long as the activation does not exceed the precision of the features. However, if the input has a large dimension  $n$ , even weak perturbations can cause large errors in the classifier [3].

There are other hypotheses that suggest that adversarial examples also arise in neural networks despite their non-linear nature, but there is very little convincing evidence for them. In fact, many commonly used

neural network architectures such as long short-term memory networks [24], maxout networks [25], and rectified linear units [26] intentionally behave linearly in order to reduce computational complexity, increasing its vulnerability to adversarial examples. However, this does not preclude non-linear architectures of neural networks such as sigmoid networks to suffer from similar problems. Many of these types of networks work mostly within a linear regime to better optimize their results, thus also leaving them open to adversarial examples as well.

While the theoretical underpinning of deep neural networks is improving, the robustness guarantees of the defence mechanisms were primarily supported by empirical observations, occasionally masking implicit assumptions. In turn, this leads to an arms race of devising defences, which are very soon bypassed by more elaborate attacks [13].

Certain models are naturally immune to adversarial examples: kernel methods, for instance, provide an example of models with this type of immunity [3, 27]. Immunity does not come for free, since kernel methods are not competitive at all with deep neural networks in terms of performance on clean datasets. This seems to suggest that performance and robustness against attacks are two important features of machine learning algorithms that are at odds, and that when developing an application based on neural networks one must choose between making it perform well or making it secure. This idea was formally expressed and proved in [28], where it is shown that robustness to adversarial attacks can be achieved by employing models of higher complexity, provided that there are no bounds on the access to new datapoints so the model underlying the data can be learned with high fidelity. The general case of a learning scheme is, however, markedly different, as the learner is typically provided with a dataset of fixed size. In the case of fixed datasets, the use of more complex models leads to overfitting, which is itself a source of weakness to adversarial attacks, and models with high capacity that fail to be as robust as models with a lower capacity are necessarily overfitting.

Methods other than neural networks may provide an advantage in this case. Bayesian inference methods have been shown to identify deviations in the uncertainty about predictions between adversarial and clean examples. This was demonstrated with Gaussian processes [18, 19], and in fact, statistical tests alone can already indicate whether samples were drawn from the same distribution as the original data [29]. Furthermore, Gaussian processes have a direct correspondence between deep and wide neural networks [30], allowing for a potentially high accuracy on clean examples, addressing the main shortcoming of kernel methods.

It still remains unclear whether other PGMs, and in particular deep variants, are more robust against adversarial examples. Many of these models are quintessentially generative, including the well-known family of Boltzmann machines. In backpropagated architectures, generative models are not necessarily robust against adversarial perturbations [3], but generative probabilistic models are yet to be benchmarked.

### 3. Minimizing free energy

While PGMs can still use some variant of gradient descent for training, the error is not backpropagated in the sense of feedforward neural networks. In fact, training involves a so-called negative phase (also known as sleep cycle or thermal equilibration) that is global in the sense that it possibly factors in everything the network has previously seen, adding some stochastic variations, and it also involves global, long-range connections. Therefore, we rightfully expect that PGMs are robust against attacks enabled by the kind of local manipulations associated to backpropagation.

Looking at it from a different angle, backpropagation with stochastic gradient descent aims at obtaining a good local optimum, whereas the minimization of free energy, which is the goal when training energy-based PGMs, gives an average over all local optima and the global optima. While an attack can easily nudge backpropagation to a different local optima, the free energy already considers all local optima, and hence its expected robustness towards perturbations.

In this work we focus on a particular type of PGMs, known as RBMs. The training of RBMs is computationally efficient with heuristics such as contrastive divergence (CD) and persistent contrastive divergence (PCD). Formally, the function to minimize when training RBMs is the negative log-likelihood,

$$\mathcal{L}_\theta(\mathcal{T}) = -\frac{1}{|\mathcal{T}|} \sum_{\mathbf{x}^{(i)} \in \mathcal{T}} \log(p_\theta(\mathbf{x}^{(i)})), \quad (2)$$

where  $\mathcal{T}$  is the training set and  $\theta$  are the parameters of the model. Intuitively, one intends to maximize the product of the probabilities of the instances in the training set, which should appear frequently (i.e. with higher probability) since they represent ‘good’ configurations of the visible units. In fact, the ideal function to minimize is the Kullback–Leibler (KL) divergence between the probability distribution that the RBM

learns,  $p_\theta(\mathbf{x})$ , and the real probability distribution over the instances  $p(\mathbf{x})$ . Discarding the term that does not depend of the model parameters from the KL divergence and assuming an approximately equal probability for good configurations to appear, equation (2) is recovered.

The parameter update rule is computed by calculating the derivatives of equation (2) with respect to  $\theta$ . For simplicity, let us choose a Boltzmann distribution in a system with hidden neurons, namely  $p_\theta(\mathbf{x}) = Z_\theta^{-1} \sum_{\mathbf{h}} \exp[-E_\theta(\mathbf{x}, \mathbf{h})]$  where the sum over all possible configurations of the units defines the partition function,  $Z_\theta = \sum_{\mathbf{x}, \mathbf{h}} \exp[-E_\theta(\mathbf{x}, \mathbf{h})]$ . Moreover, let us define the free energy of a configuration of visible units  $\mathcal{F}_\theta(\mathbf{x})$  by

$$e^{-\mathcal{F}_\theta(\mathbf{x})} = \sum_{\mathbf{h}} e^{-E_\theta(\mathbf{x}, \mathbf{h})}. \quad (3)$$

Using this, the loss function can be rewritten as

$$\mathcal{L}_\theta(\mathcal{T}) = \log \sum_{\mathbf{x}} e^{-\mathcal{F}_\theta(\mathbf{x})} + \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x}^{(i)} \in \mathcal{T}} \mathcal{F}_\theta(\mathbf{x}^{(i)}). \quad (4)$$

The second sum is over the training set, while the first one is over *all* possible neuron configurations. This is the term whose computation is intractable classically—for a model with  $N$  binary neurons, the total number of possible configurations is  $2^N$ . Now, computing the gradient of equation (4), we observe two different terms:

$$\frac{\partial \mathcal{L}_\theta}{\partial \theta}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x}^{(i)} \in \mathcal{T}} \frac{\partial \mathcal{F}_\theta(\mathbf{x}^{(i)})}{\partial \theta} - \sum_{\mathbf{x}} \left[ p_\theta(\mathbf{x}) \frac{\partial \mathcal{F}_\theta(\mathbf{x})}{\partial \theta} \right]. \quad (5)$$

Here we identify the *positive* phase as the term coming from the evaluations on the training set, and the *negative* phase as that coming from evaluations on all possible configurations of visible units. The simplest energy function associated to an RBM is given by

$$E_\theta(\mathbf{x}, \mathbf{h}) = - \sum_i x_i b_i - \sum_j h_j c_j - \sum_{ij} w_{ij} x_i h_j, \quad (6)$$

where  $\theta$  is now  $\{w_{ij}, b_i, c_j\}_{ij}$ . The free energy can be thus expressed as

$$\mathcal{F}_\theta(\mathbf{x}) = - \sum_i x_i b_i - \sum_j \log \left( 1 + e^{c_j + \sum_i w_{ij} x_i} \right). \quad (7)$$

Picking any derivative, we can see a more explicit formulation of the positive and negative phases. For instance, if we take the visible biases,  $b_i$ , with equation (5) we obtain

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial b_i}(\mathcal{T}) &= \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x}^{(k)} \in \mathcal{T}} (-x_i^{(k)}) - \sum_{\mathbf{x}} p_\theta(\mathbf{x}) (-x_i) \\ &= -(\langle x_i \rangle_{\text{data}} - \langle x_i \rangle_{\text{model}}). \end{aligned} \quad (8)$$

The positive phase,  $\langle \cdot \rangle_{\text{data}}$ , is easy to compute since it only involves averages over the training set, but the negative phase,  $\langle \cdot \rangle_{\text{model}}$ , is not possible to calculate explicitly since it requires knowledge of the overall distribution  $p_\theta(\mathbf{x})$ , or analogously of the partition function  $Z_\theta$ .

### 3.1. Classical heuristics

To evaluate the gradient correctly one needs to be able to satisfactorily approximate the negative phase,  $\langle \cdot \rangle_{\text{model}}$ . A way of obtaining samples from the model distribution is via Markov chain Monte Carlo Gibbs sampling [31]: starting from a random configuration of the visible neurons, one iterates sampling the hidden and visible neurons from the corresponding conditional distributions until the stationary state is reached. This proper estimation is prohibitively expensive due to the large amount of iterations required to reach the stationary state, but some simple heuristics work decently in practice.

CD is one of such simple heuristics [32]. Instead of iterating infinitely over the layers of the RBM to find  $\langle \cdot \rangle_{\text{model}}$ , when using CD- $k$  one performs  $k$  iterations of a batch of  $n$  chains that are initialized in training datapoints, and computes the expectation as averages over those  $n$  chains. This is an extremely simplified way of doing Gibbs sampling without a burn-in time, and instead of starting from random initialization, one starts from a clamped set of visible nodes. Choosing to start from training data in the negative phase has two benefits: (i) the resultant gradient will produce a model that does not drift too far away from the ground

truth, and (ii) there is no need to wait for the Markov chain to mix in order to draw samples of reasonable quality.

A more reliable variant of CD is called PCD [33], where instead of initializing the chains for every average to be computed, the initial seeds for a sampling process are the final state of the chains in the previous process. This may reduce the impact of burn-in since a Markov chain is constantly maintained. However, one has to carefully balance between the mixing rate and weight updates [34].

While seemingly useful in practice, one must bear in mind that by using these heuristics one may lose visibility of discovering samples that follow more closely the Boltzmann distribution. Moreover, these sampling heuristics obviate that, especially in the early stages of training a Boltzmann machine, the fact that the weights are initialized randomly makes the proper characterization of the low-energy sector an NP-complete problem [35, 36].

### 3.2. Quantum-enhanced sampling

A powerful realization that has arisen with the recent advent of analog quantum computers is that many algorithms which had an inspiration in the dynamics of physical processes and that simulated or approximated those physical processes can now be readily implemented in analog computers, without approximations. In the context of the NP-complete problem of computing the negative phase of a Boltzmann machine, many state-of-the-art methods have inspiration from physics, such as simulated annealing [37] or parallel tempering [38]. Now it is possible not just to simulate these dynamics, but to prepare a physical system in the particular desired state and measure or sample it directly in order to compute  $\langle \cdot \rangle_{\text{model}}$ . It is well known that more accurate approximations of the negative phase give rise to improvements in terms of model quality and training time [33, 36]. Thus, it is reasonable to expect that the direct access to model samples from measuring a physical system results in higher-quality models that, from the perspective of adversarial attacks, are more robust than models that are trained using classical heuristics. Alternatively, one can reasonably expect to match the robustness of models trained using classical heuristics in fewer training cycles by using analog sampling methods.

In the early training stage of a Boltzmann machine, the Markov chains used in many heuristics for approximating the log-likelihood gradient may behave as expected because the weights are usually initialized as small values, which means that the Markov chains just need to approximate samples drawn from distributions close to thermal noise. At later stages, however, it becomes harder to keep the chains near their stationary distributions [39]. It is now possible, however, to prepare systems in the stationary distribution directly, instead of approximating it. The D-Wave quantum annealer is a physical device that has a working mechanism similar to simulated annealing [40]. In addition to using the thermal fluctuations inherent to any physical system, quantum annealing (QA) further explores the search space by exploiting quantum fluctuations. Studies have shown that large-scale quantum annealers such as those exemplified by the D-Wave quantum computers can offer significant speedups for certain problem classes [41].

Reference [42] conducted extensive experiments that compare the sampling performance of QA against CD. For energy functions with high barriers QA is more effective in exploring complex search spaces, which leads to a more accurate calculation of gradients. Computational efficiency aside, CD and PCD may not yield high quality models that are competitive with backpropagated architectures in terms of prediction accuracy, despite the fact that they might be more robust to attacks. This has inspired further heuristics such as Boltzmann-Encoded Adversarial Machines [43], where a GAN-like setting is combined with an RBM architecture. Nevertheless, CD and PCD are just approximations of thermal sampling training schemes, which can be performed exactly by hardware-based sampling to yield intrinsically high-quality Boltzmann machines. In particular, quantum-enhanced sampling on contemporary quantum annealers has already shown notable advantages against CD and PCD training [44, 45].

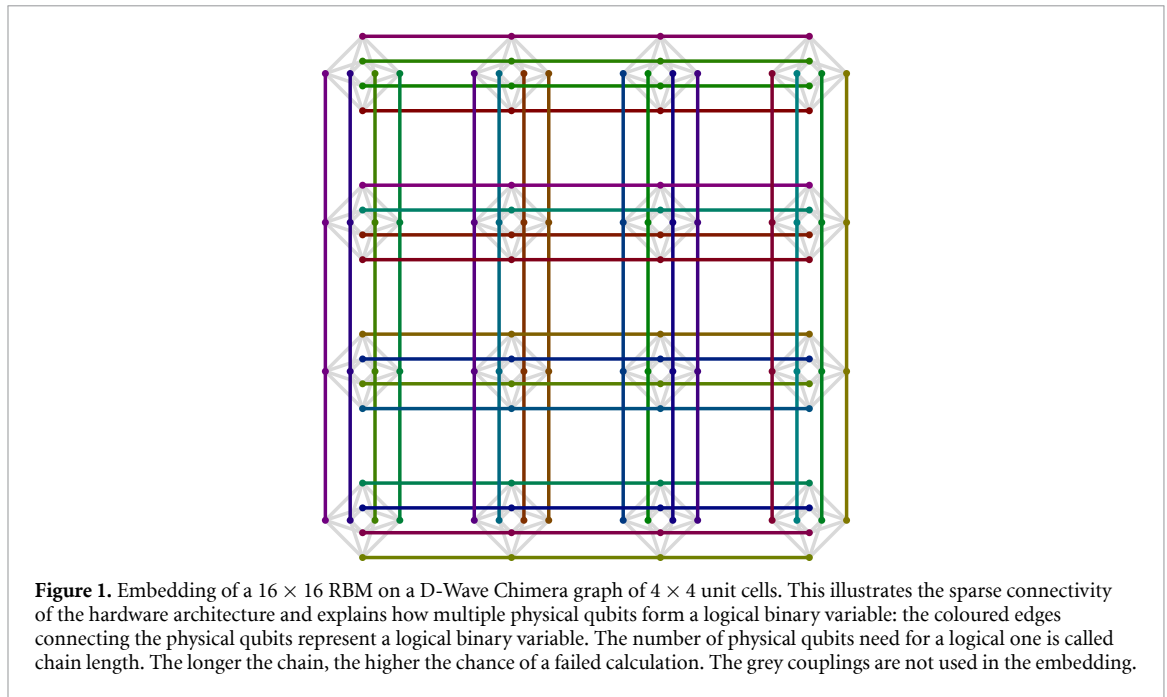
When working with QA one must make a distinction between two different distributions. The first one is the actual Boltzmann distribution with energy function defined in equation (6). The second one is the quantum annealer sampling distribution. In theory, once the binary units are replaced by qubits as in QA, the energy function (6) is substituted by the Hamiltonian [46]

$$H_{\theta}(\sigma_x^z, \sigma_h^z) = \sum_x b_x \sigma_x^z + \sum_h c_h \sigma_h^z + \sum_{x,h} w_{xh} \sigma_x^z \sigma_h^z, \quad (9)$$

where  $\sigma_i^z$  represents a spin on a lattice site  $i$  that takes a value in  $\{+1, -1\}$ . Technically, these spins can be in a superposition, which means that they are equivalent to qubits. Equation (6) will represent the diagonal elements of the  $2^N \times 2^N$  Hamiltonian matrix described by equation (9), and  $N$  is the total amount of neurons of the Boltzmann machine. The partition function then becomes the trace of  $e^{-H_{\theta}}$ , i.e.  $Z_{\theta} = \text{Tr}[e^{-H_{\theta}}]$ .

Unlike classical Boltzmann machines where approximations have to be made to circumvent the evaluation of the partition function, in a quantum annealer one can prepare physical samples quantum





mechanically, without the need of any explicit calculation. Instead, the samples drawn from the annealer directly follow the desired Boltzmann distribution.

Despite its sampling capacity, qubits on the D-Wave quantum computer are sparsely connected, which means that not any two qubits can be connected (or coupled) by a  $w$ . In the current generation of the D-Wave chip, namely D-Wave 2000Q, each qubit is connected to a maximum of six qubits. In order to couple qubits that are far apart one needs to resort to embeddings. An embedding requires building virtual qubits by chaining up multiple physical qubits with strong ferromagnetic couplings. Shorter chains are favoured against longer ones, since in general, the longer the chain, the more likely it will break, which would distort the result of the sampling. In figure 1 we illustrate the embedding of a  $16 \times 16$  RBM on a  $4 \times 4$  D-Wave Chimera graph, where each unit in the RBM is composed of four connected qubits on the quantum chip, as shown by the coloured chains. In this work, we use the heuristic *minor-embedding* method to map the graphical structure of the RBM architectures employed to the D-Wave Chimera structure [47].

Another important issue one needs to address when performing QA is the so-called *temperature effect*. There are two factors to the temperature effect. First, the programmable parameters on the D-Wave chip are normally constrained within a specific range. When an RBM is being trained, one needs to scale the model to make sure it fits within the parameter range on the chip. Scaling the energy function of equation (6) can be considered as changing the temperature of the problem. Second, the chip runs at a very low temperature—around the milli-Kelvin level—which implies that when the Hamiltonian of equation (9) is solved on the D-Wave quantum annealer, it is actually solved with a scaling factor of  $T \ll 1$ . The combination of these two factors will ultimately determine the temperature at which the samples are returned.

#### 4. Experimental setup

We outline now the different procedures for attacking and defending classifier algorithms, and detail the parameter choices for the classical and quantum RBMs.

The general attack-defence procedure in neural networks is as follows: first, each of the defences is applied to the original dataset, producing ‘defended’ training and test datasets. Then, each ‘defended’ training dataset is employed to finetune a ResNet18 convolutional neural network (CNN), and each ‘defended’ test set is used to build datasets of adversarial images using each of the attacks. The attack datasets are 1:1 with the test datasets, so every image in each defended dataset has an adversarial version in each of the attack datasets. Finally, the robustness to adversarial attacks is measured as the percentage of test adversarial images that are misclassified by the CNN.

The choice of attacks and defences were chosen to reflect methods that are commonly used and that proved strong in previous benchmarks [17]. While the defences make use of no knowledge about the classifier they will defend (i.e. they are black-box), we consider white-box attacks where the adversary has access to the classifier’s architecture and parameters. Black-box attacks, which only have access to queries of

the model, are in principle more ‘dangerous’ in that they can easily apply to a wide variety of classifiers. Prior work has proven, however, that it is possible for an attacker to train a model on black-box queries of the target, and use this white-box substitute model to build attacks that are effective in the target model [12]. Thus, considering just white-box attacks is realistic and sufficient. Finally, all the attacks are non-targeted: the focus is put in creating a misclassification, rather than forcing a particular label to be predicted.

As for the proposal of this work (recall, replacing the CNN classifier by an RBM generative model that is later repurposed into a classifier), the defence consists of simply training the RBM on the original dataset. The set of adversarial attacks are built from the original dataset as well.

## 4.1. Attacks

### 4.1.1. FGSM

FGSM is a standard attack for generating adversarial images. The method has been described in detail in the Related Work section.

### 4.1.2. Carlini and Wagner

Carlini and Wagner proposed gradient-based attacks that have been found to be among the most effective when using small perturbations [9]. It re-frames generating adversarial examples as an optimization problem. Clearly, this is a difficult problem to solve and multiple techniques are employed to simplify the optimization. Firstly, a binary search algorithm is employed to find a suitable coefficient for the optimization terms, and after that, the optimization terms are converted to *arctanh* space so that efficient, state-of-the-art optimization solvers can be utilized.

### 4.1.3. Deepfool

This method uses a different insight to derive efficient attacks [5]. It characterizes robustness as the distance between a data point and the decision surface. The aim of the adversary is to minimize the perturbation in the  $L_2$  norm while misleading the classifier. DeepFool represents the faces of a polyhedron with the decision boundary planes from the classifier to describe the output space. The attack then finds the minimal perturbation that changes the classifier’s decision within the polyhedron.

## 4.2. Baseline defences

### 4.2.1. Adversarial training

The most popular defence is plain adversarial training, in the spirit of GANs [2]. In contrast with the use of GANs for generative purposes, in the case of adversarial training it is the discriminator, and not the generator, that which one is interested in improving its performance. This method injects adversarial examples from the attacks during the training phase, so the discriminator is trained either on a mix of both clean and adversarial examples, or on only the latter. However, since the adversarial examples are generated with a specific type of attack, the discriminator remains vulnerable to other types of attacks.

Adversarial effects can also be efficiently mitigated through randomization, effectively ending up to a form of data augmentation [48]. Techniques commonly used include random resizing and random padding. The rationale behind these defences is that iterative attacks could overfit specific network parameters, hence low-level image transformations can destroy the structure of adversarial perturbations.

This defence is of a nature different than the remaining, and thus is implemented in a different way. Instead of applying the defence to the training dataset and training the CNN in it, in adversarial training the CNN is trained directly in the union of the original clean dataset and the result of the application of the attack to the original clean dataset.

### 4.2.2. Feature squeezing

Feature squeezing [49] is a computationally inexpensive yet powerful state-of-the-art method for defending against adversarial attacks. This method reduces the colour bit depth of each pixel in the images used for training. When representing images, colour bit depths are often employed to display images that are very close to their natural counterpart. However, the features that are created as a result of this colour bit depth are often not necessary for recognizing what an image is representing. Therefore, reducing the colour bit depth can in theory also reduce the opportunities for an attack to find an adversarial image without sacrificing classifier accuracy.

### 4.2.3. Spatial smoothing

Local spatial smoothing [49], which is related to feature squeezing, is a method consisting of reducing image noise. It can be thought of as a sliding window that centres around each pixel in the image and replaces the pixel with the median value of each of the neighbouring pixels. This creates a ‘blur’ over the image and helps



mitigate the effects of adversarial perturbations (especially when applied to salt-and-pepper noise), while maintaining the features that make correct classification possible.

#### 4.2.4. Sampling-based defences

The method we propose, and furthermore we show effective for defending against adversarial attacks in the next section, is abandoning the concept of training a discriminative neural network and instead training a generative model on the joint dataset of inputs and corresponding targets. This is, instead of parametrizing a function  $f_\theta(x)$  by a neural network architecture and using backpropagation to minimize some notion of distance between the prediction,  $f_\theta(x)$ , and the corresponding true label,  $y$ , we propose learning the joint distribution of datapoints and corresponding labels,  $p_\theta(x, y)$ —provided as the Boltzmann distribution of some parametrized energy function—, by optimizing the negative log-likelihood given in equation (2) where the input  $\mathbf{x}$  is given as the concatenation of a datapoint  $x$  and its corresponding label  $y$ . Once the model is trained, the label assigned to a datapoint is that which minimizes the free energy of equation (7) when the neurons encoding the datapoint are fixed. In the case of a small possible choice of labels, this can be determined by direct computation of the free energy of all the neuron configurations  $\mathbf{x}_\ell = (x, \ell) \forall \ell = 1, \dots, |\mathcal{Y}|$  that correspond to the concatenation of the datapoint and a valid label, and choosing the label for which the free energy is lowest, this is,

$$y(x) = \operatorname{argmin}_\ell \mathcal{F}(\mathbf{x}_\ell) \text{ s.t. } \mathbf{x}_\ell = (x, \ell). \quad (10)$$

In more complicated cases, one can resort to sampling the label from the conditional distribution  $p_\theta(y|x)$  via Markov Chain Monte Carlo methods.

In the particular case of this work we focus on RBM models, where the computation of the positive phase of the gradient update is easy. We implement them using the *ebm-torch* package [50], setting the number of hidden neurons to 40, the learning rate to 0.01, and using the Adam algorithm for updating the parameters in the graph. No dropout or other regularization methods are used.

We use two different techniques for estimating the negative phase. On one hand we do a classical training, using PCD to obtain the approximations of  $\langle \cdot \rangle_{\text{model}}$ . We choose a small batch size of 10 and long iteration chains ( $k = 50$ ). On the other, we compute it directly from the physical sampling of a thermal quantum state of a set of qubits in the D-Wave 2000Q processor. In this case we use a larger batch size of 1000 for computing the positive phases, which is a more natural choice for the hardware. In light of the temperature effect discussed in section 3.2, we rescale the coefficients manually to have a tight control over the temperature at which the annealing is done. We request 500 samples from the chip, which are classically postprocessed by the D-Wave server-side software stack. Each quantum sample is obtained with a 20  $\mu\text{s}$  annealing time. To mitigate the intrinsic control errors that might exist on the quantum chip, we randomly choose five spin reversal transformations, which means that the Hamiltonian remains unchanged if, along with the random flipping of spins, the signs of bias and coupling terms are flipped accordingly. Since we can quickly reach a reasonably good model with heuristic methods, we bootstrap the quantum-enhanced RBMs with a partially trained classical RBM pre-trained over 100 epochs and leave the heavy lifting to the quantum annealer at a later training stage.

## 5. Evaluation

The Modified National Institute of Standards and Technology (MNIST) dataset provides a simple and robust baseline that is useful to understand some limitations of defence mechanisms [51]. While arguably it should not be the only benchmark [13], we rely on it exclusively because this is the only commonly used image dataset that can easily be transformed to a current quantum processing unit. A larger image resolution or additional colour channels would require an extensive transformation pipeline or the addition of convolutional layers before the RBM, which could introduce confounding effects in the evaluation.

Tables 1 and 2 summarize the results. We used two versions of MNIST: the original one, containing  $28 \times 28$  greyscale images, and a rescaled one, containing  $7 \times 7$  binary images. The latter transformation was necessary to embed the RBM in the quantum annealer, as the original version would require a significantly higher number of qubits and connectivity than what is available with contemporary technology.

The columns of the tables refer to the defences. The RBM is the classical version trained with PCD, whereas the QRBM in table 2 is the version trained using the D-Wave quantum annealer. The attacks are all implemented using the *foolbox* library [52].

RBM proves significantly more robust against attacks even when classical sampling heuristics are used during training. This underlies the importance of architectures that have an internal equilibration step, as opposed to purely greedy backpropagation of errors. This benefit of using the RBM can be tracked back to

**Table 1.** Accuracy of defences after attacks on the original MNIST test dataset (10 000  $28 \times 28$  greyscale images of handwritten digits).

	Adv. training (%)	Feat. squeezing (%)	Spatial smoothing (%)	RBM (%)
FGSM	13.39	20.24	18.54	<b>81.18</b>
DeepFool	12.80	19.10	19.70	<b>83.36</b>
CarliniWagner	72.00	78.54	45.80	<b>83.16</b>

Note: The values in bold highlight the best-performing defence for each attack.

**Table 2.** Accuracy of defences after attacks on the MNIST test dataset downsampled to  $7 \times 7$  and binarized. This transformation is necessary for training the RBM on the quantum annealer.

	Adv. training (%)	Feat. squeezing (%)	Spatial smoothing (%)	RBM (%)	QRBM (%)
FGSM	19.36	28.03	14.74	<b>87.15</b>	84.49
DeepFool	21.03	26.64	13.66	84.50	<b>85.18</b>
CarliniWagner	48.24	77.15	22.00	<b>85.71</b>	82.92

Note: The values in bold highlight the best-performing defence for each attack.

the generative nature of the model. Instead of learning a ‘simple’ class-conditional density like discriminative models, the RBM seeks to approximate the entire data distribution. While this is a more difficult problem that may result in a lower performance on an unperturbed dataset, it also reduces the overfitting of the model, making it more capable of generalizing to adversarial images. Referring back to [28], this result is an extension of the upper bound of the loss function for more complex models. Clearly, complex models drawn from a rich class of hypotheses must have loss functions smaller than simpler models, given (and this is an important point) that enough training data is provided. While the CNNs on which the baseline defences were established are much more complex than an RBM, their performance against adversarial examples demonstrates that they overfit to the training and test data.

Importantly, quantum-enhanced sampling is on par with the classical RBM results and beats backpropagation-based architectures. While it only outperforms classical PCD in only one attack and by a small margin, given the current immature state of quantum technologies, this is already a remarkable result. Furthermore, it highlights an important aspect of the use of quantum technologies in machine learning which is more directly applicable and easier to measure than the speedup claims based in computational complexity arguments [53].

## 6. Conclusions and discussion

We have shown that using generative PGMs as discriminators produces models that are much more robust to adversarial attacks than standard discriminators based on deep neural networks. RBMs trained on the MNIST dataset are notably more robust to misclassification of crafted images designed to fool the classification than standard CNNs, even when defensive pre-processings of the datasets are applied in order to protect the training pipeline. Moreover, using quantum-enhanced estimations of the log-likelihood gradient of the RBM resulted in comparable improvements, having one instance, the DeepFool attack, where the quantum-enhanced RBM was more robust than its classical counterpart.

The source of the robustness of generative PGMs to adversarial attacks can be ascribed to the overfitting to training (and available test) data that discriminative models present when trained on finite datasets. However, given that the primary goal of the generative model is not the classification accuracy but the efficient approximation of the data manifold, its utility is lower than the achievable with discriminative models. This is an expression of the no-free-lunch results of [28] applied to training on finite datasets.

Further support of the claim that the non-discriminative learning procedure of PGMs is a major factor in the robustness of the model can be found in [54]. It is observed that human learning relies mostly on a highly structured learning mechanism and unsupervised learning rather than supervised. Since humans are much better than current machine learning techniques at defending against adversarial examples, mimicking how we learn may be the best way to design robust defences. Moreover, graphical models in general are structured in a much more nuanced way than neural networks. The structure of most graphical models encodes some prior knowledge about the nature of the system it is modelling, something that seems to be also present in the human brain. The similarities of graphical models to the way humans learn not only points to a reason why RBMs may be more robust to adversarial examples than neural networks, but also suggests that other, more structured graphical models may be even more effective.

From the quantum perspective, quantum computers are often touted for a significant speedup expressed in terms of computational complexity or runtime in some specific setting. However, the current mismatch

between theoretical requirements for provable speedups and experimental feasibility motivates the exploration of alternative metrics by which quantum computing may provide advantages [55]. Here we shift the attention from a speedup to a qualitative difference that quantum computers could offer in training energy-based PGMs. It is known that the quality of the trained model increases and the number of training epochs decreases with better approximations of the negative phase of the gradient (a simple example is shown in [36]). In the context of robustness to adversarial attacks, it is thus expected that the improvements in training enabled by the direct access to samples drawn from the model distribution (via directly measuring the qubits in the D-Wave quantum annealer) provides more robust models using fewer computational resources. The results that we present do not yet allow to make a strong claim regarding an advantage, but they signal to promising a research direction where quantum advantages may be easier to achieve in the current era of noisy, intermediate-scale quantum computers.

Although we are encouraged by the robustness of Boltzmann machines trained with QA to adversarial attacks demonstrated at this stage (which is, recall, comparable to those of models trained using classical heuristics), the size of the available quantum chips is not yet scalable to industry-standard datasets. This is not only due to the number of qubits, but also to their sparse connectivity. In fact, the sparse connectivity requires that, currently, most models of interest must be embedded in the chip's architecture by using additional qubits. Available embedding heuristics are designed to preserve the ground state, but access to the low-energy excited states is also needed when training Boltzmann machines. An important direction of subsequent research, that does not necessarily require experimental advances, is thus understanding the relations between the energy spectra of the original problem and its embedding in the chip.

In future work, we plan to extend the current analysis beyond RBMs. By using a richer class of models—like fully connected Boltzmann machines—one can expect that its larger expressibility allows to improve the performance of classification while maintaining the adversarial robustness. However, even simple modifications like lateral connections being introduced into the hidden layer require hardware sampling—or alternative approximations—to be used also when computing the positive phase. Even if one can extend the range of Boltzmann machines that can be trained, the number of qubits needs a dramatic increase for performing an end-to-end training in the quantum processing units. As [45] pointed out, classical preprocessing is recommended. A natural way to extend this work is to include convolutional layers and attach an RBM to the final layer.

We considered just white-box attacks since these can also be built when one only has black-box access to the target model [12]. The application of the attacks to classical RBMs is straightforward. For the quantum RBM, in contrast, white-box access (i.e. the access not to the parameters that are input to the annealer, but those characterizing the Boltzmann distribution of the samples) requires extensive calls to the quantum annealer, which is prohibitive at the moment, particularly if we factor in the tuning requirements of both the hardware and the Boltzmann machines. Nevertheless, as the software stack improves and the cost of access drops, we expect that the white-box scenario will eventually become viable.

Other than the D-Wave quantum annealer, there are currently further choices of analog sampling platforms. For example, degenerate optical parametric oscillators can simulate an Ising model, with the desired belief that they do not suffer from sparse connectivity [56]. Furthermore, fully classical application-specific integrated circuits are being tested for digital annealing [57]. It is interesting to see if the robustness achieved with the D-Wave quantum annealer applies to other hardware technologies, which would indicate that the improvements are indeed consequence of the use of PGMs in classification, rather than of some other feature of QA.

## Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

## Acknowledgments

This work was done before Peter Wittek left us, and is dedicated to his memory. Peter was a great leader and mentor, which had great impact not only in the authors' careers, but in that of so many others. Peter, you would have enjoyed so much the advances we are witnessing. The QML community still misses you, and will keep doing so for a long, long time. A P-K is grateful to the Creative Destruction Lab for their hospitality. A P-K's work was supported by Fundació Obra Social 'la Caixa' (LCF/BQ/ES15/10360001) and the European Union's Horizon 2020 research and innovation programme—Grant Agreement No. 648913.

## ORCID iDs

Yanbo Xue  <https://orcid.org/0000-0001-5999-1521>

Alejandro Pozas-Kerstjens  <https://orcid.org/0000-0002-3853-3545>

## References

- [1] Su J, Vasconcellos Vargas D and Kouichi S 2019 One pixel attack for fooling deep neural networks *IEEE Trans. Evol. Comput.* **23** 828
- [2] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R 2014 Intriguing properties of neural networks *Int. Conf. on Learning Representations* (arXiv:1312.6199)
- [3] Goodfellow I J, Shlens J and Szegedy C 2015 Explaining and harnessing adversarial examples *Int. Conf. on Learning Representations* (arXiv:1412.6572)
- [4] Cubuk E D, Zoph B, Schoenholz S S and Le Q V 2018 Intriguing properties of adversarial examples *Int. Conf. on Learning Representations* (arXiv:1711.02846)
- [5] Moosavi-Dezfooli S-M, Fawzi A and Frossard P 2016 DeepFool: a simple and accurate method to fool deep neural networks 2016 *Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 2574–82
- [6] Šrندیć N and Laskov P 2014 Practical evasion of a learning-based classifier: a case study *Proc. SP-4, Symp. on Security and Privacy* pp 197–221
- [7] Grosse K, Papernot N, Manoharan P, Backes M and McDaniel P 2017 Adversarial examples for malware detection *Proc. ESORICS-17, Symp. on Research in Computer Security* (Berlin: Springer) pp 62–79
- [8] Cisse M M, Adi Y, Neverova N and Keshet J 2017 Houdini: Fooling deep structured visual and speech recognition models with adversarial examples *Advances in Neural Information Processing Systems* vol 30, ed I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (Red Hook, NY: Curran Associates, Inc.) pp 6977–87 (arXiv:1707.05373)
- [9] Carlini N and Wagner D 2017 Towards evaluating the robustness of neural networks *Proc. Symp. on Security and Privacy* pp 39–57
- [10] Brendel W, Rauber J and Bethge M 2018 Decision-based adversarial attacks: reliable attacks against black-box machine learning models *Int. Conf. on Learning Representations* (arXiv:1712.04248)
- [11] Zhao Z, Dua D and Singh S 2018 Generating natural adversarial examples *Int. Conf. on Learning Representations* (arXiv:1710.11342)
- [12] Papernot N, McDaniel P and Goodfellow I 2016 Transferability in machine learning: from phenomena to black-box attacks using adversarial samples (arXiv:1605.07277)
- [13] Carlini N and Wagner D 2017b Adversarial examples are not easily detected: bypassing ten detection methods *Proc. 10th ACM Workshop on Artificial Intelligence and Security AISec '17* (New York: Association for Computing Machinery) pp 3–14
- [14] Carlini N and Wagner D 2017c MagNet and “Efficient defenses against adversarial attacks” are not robust to adversarial examples (arXiv:1711.08478)
- [15] Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A 2018 Towards deep learning models resistant to adversarial attacks *Int. Conf. on Learning Representations* (arXiv:1706.06083)
- [16] Fawzi A, Fawzi H and Fawzi O 2018 Adversarial vulnerability for any classifier *Proc. 32nd Int. Conf. on Neural Information Processing Systems NIPS'18* (Red Hook: Curran Associates Inc.) pp 1186–95 (arXiv:1802.08686)
- [17] Kurakin A et al 2018 Adversarial attacks and defences competition (arXiv:1804.00097)
- [18] Bradshaw J, Matthews A G de G, and Ghahramani Z 2017 Adversarial examples, uncertainty, and transfer testing robustness in Gaussian process hybrid deep networks (arXiv:1707.02476)
- [19] Grosse K, Pfaff D, Smith M T and Backes M 2017b How wrong am I? —studying adversarial examples and their impact on uncertainty in Gaussian process machine learning models (arXiv:1711.06598)
- [20] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* vol 27, ed Z Ghahramani, M Welling, C Cortes, N Lawrence and K Q Weinberger (Red Hook: Curran Associates, Inc.) pp 2672–80 (arXiv:1406.2661)
- [21] Arute F et al 2019 Quantum supremacy using a programmable superconducting processor *Nature* **574** 505
- [22] Zhong H-S et al 2020 Quantum computational advantage using photons *Science* **370** 1460–63
- [23] Yuan X, He P, Zhu Q and Li X 2019 Adversarial examples: attacks and defenses for deep learning *IEEE Trans. Neural Netw. Learn. Syst.* **30** 2805–24
- [24] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80
- [25] Goodfellow I, Warde-Farley D, Mirza M, Courville A and Bengio Y 2013 Maxout networks *Proc. 30th Int. Conf. on Machine Learning Proc. of Machine Learning Research* vol 28, ed S Dasgupta and D McAllester (Atlanta: PMLR) pp 1319–27 (arXiv:1302.4389)
- [26] Jarrett K, Kavukcuoglu K, Ranzato M and LeCun Y 2009 What is the best multi-stage architecture for object recognition? *Proc. ICCV-09, 12th Int. Conf. on Computer Vision* pp 2146–53
- [27] Hein M and Andriushchenko M 2017 Formal guarantees on the robustness of a classifier against adversarial manipulation *Advances in Neural Information Processing Systems* vol 30, ed I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (Red Hook: Curran Associates, Inc.) pp 2266–76 (arXiv:1705.08475)
- [28] Papernot N, McDaniel P, Sinha A and Wellman M 2018 SoK: security and privacy in machine learning 2018 *IEEE Symp. on Security and Privacy (EuroSP)* pp 399–414
- [29] Grosse K, Manoharan P, Papernot N, Backes M and McDaniel P 2017c On the (statistical) detection of adversarial examples (arXiv:1702.06280)
- [30] Lee J, Bahri Y, Novak R, Schoenholz S S, Pennington J and Sohl-Dickstein J 2018 Deep neural networks as Gaussian processes *Int. Conf. on Learning Representations* (arXiv:1711.00165)
- [31] Murphy K P 2012 *Machine Learning: A Probabilistic Perspective* (Cambridge, MA: MIT Press)
- [32] Sutskever I and Tieleman T 2010 On the convergence properties of contrastive divergence *Proc. Thirteenth Int. Conf. on Artificial Intelligence and Statistics* pp 789–95 (<http://proceedings.mlr.press/v9/sutskever10a.html>)
- [33] Tieleman T 2008 Training restricted Boltzmann machines using approximations to the likelihood gradient *Proc. 25th Int. Conf. on Machine Learning* pp 1064–71
- [34] Tieleman T and Hinton G 2009 Using fast weights to improve persistent contrastive divergence *Proc. 26th Annual Int. Conf. on Machine Learning* (New York: Association for Computing Machinery) pp 1033–40
- [35] Barahona F 1982 On the computational complexity of Ising spin glass models *J. Phys. A: Math. Gen.* **15** 3241

- [36] Pozas-Kerstjens A, Muñoz-Gil G, Piñol E, Garcia-March M A, Acin A, Lewenstein M and Grzybowski P R 2021 Efficient training of energy-based models via spin-glass control *Mach. Learn.: Sci. Technol.* **2** 025026
- [37] Marinari E and Parisi G 1992 Simulated tempering: a new Monte Carlo scheme *EPL* **19** 451
- [38] Desjardins G, Courville A, Bengio Y, Vincent P and Delalleau O 2010 Tempered Markov chain Monte Carlo for training of restricted Boltzmann machines *Proc. Thirteenth Int. Conf. on Artificial Intelligence and Statistics* Chia Laguna Resort, Sardinia vol 9, ed Y W Teh and M Titterton (Chia Laguna Resort, Sardinia: JMLR Workshop Conf. Proc.) pp 145–52 (<https://proceedings.mlr.press/v9/desjardins10a.html>)
- [39] Salakhutdinov R and Hinton G 2012 An efficient learning procedure for deep Boltzmann machines *Neural Comput.* **24** 1967–2006
- [40] Lanting T et al 2014 Entanglement in a quantum annealing processor *Phys. Rev. X* **4** 021041
- [41] Denchev V S, Boixo S, Isakov S V, Ding N, Babbush R, Smelyanskiy V, Martinis J and Neven H 2016 What is the computational value of finite range tunneling? *Phys. Rev. X* **6** 031015
- [42] Korenkevych D, Xue Y, Bian Z, Chudak F, Macready W G, Rolfe J, and Andriyash E 2016 Benchmarking quantum hardware for training of fully visible Boltzmann machines (arXiv:1611.04528)
- [43] Fisher C K, Smith A M, and Walsh J R 2018 Boltzmann encoded adversarial machines (arXiv:1804.08682)
- [44] Benedetti M, Realpe-Gómez J, Biswas R and Perdomo-Ortiz A 2017 Quantum-assisted learning of hardware-embedded probabilistic graphical models *Phys. Rev. X* **7** 041052
- [45] Benedetti M, Realpe-Gómez J and Perdomo-Ortiz A 2018 Quantum-assisted Helmholtz machines: a quantum—classical deep learning framework for industrial datasets in near-term devices *Quantum Sci. Technol.* **3** 034007
- [46] Amin M H, Andriyash E, Rolfe J, Kulchitskiy B and Melko R 2018 Quantum Boltzmann machine *Phys. Rev. X* **8** 021050
- [47] Cai J, Macready W G and Roy A 2014 A practical heuristic for finding graph minors (arXiv:1406.2741)
- [48] Xie C, Wang J, Zhang Z, Ren Z and Yuille A 2018 Mitigating adversarial effects through randomization *Int. Conf. on Learning Representations* (arXiv:1711.01991)
- [49] Xu W, Evans D and Qi Y 2018 Feature squeezing: detecting adversarial examples in deep neural networks *25th Annual Network and Distributed System Symp. (NDSS)* (San Diego, CA: The Internet Society) pp 18–21
- [50] Pozas-Kerstjens A 2018 *ebm-torch: energy-based models in PyTorch* GitHub repository ([www.github.com/apozas/ebm-torch](http://www.github.com/apozas/ebm-torch))
- [51] Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D and McDaniel P 2018 Ensemble adversarial training: attacks and defenses *Int. Conf. on Learning Representations* (arXiv:1705.07204)
- [52] Rauber J, Brendel W and Bethge M 2017 Foolbox: A Python toolbox to benchmark the robustness of machine learning models *Reliable Machine Learning in the Wild Workshop, 34th Int. Conf. on Machine Learning* (arXiv:1707.04131)
- [53] Zhao Z, Pozas-Kerstjens A, Rebertost P and Wittek P 2019 Bayesian deep learning on a quantum computer *Quantum Mach. Intell.* **1** pp 41–51
- [54] Zador A M 2019 A critique of pure learning and what artificial neural networks can learn from animal brains *Nat. Commun.* **10** 3770
- [55] Weber M, Liu N, Li B, Zhang C and Zhao Z 2020 Optimal provable robustness of quantum classification via quantum hypothesis testing (arXiv:2009.10064)
- [56] QNNCloud 2017 Quantum neural network: optical neural networks operating at the quantum limit (<https://qnncloud.com/>)
- [57] Fujitsu 2017 Quantum computing and AI start a new era Fujitsu Journal (<http://journal.jp.fujitsu.com/en/2017/12/13/01/>)