Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Machine learning in physics: the pitfalls of poisoned training sets

To cite this article: Chao Fang *et al* 2020 *Mach. Learn.: Sci. Technol.* **1** 045001

View the article online for updates and enhancements.

This content was downloaded from IP address 122.172.173.154 on 29/06/2023 at 10:54

IOP Publishing Mach. Learn.: Sci. Technol. 1 (2020) 045001 https://doi.org/10.1088/2632-2153/aba821

MACHINE LEARNING Science and Technology

OPEN ACCESS

RECEIVED 17 March 2020

REVISED 3 July 2020

ACCEPTED FOR PUBLICATION 21 July 2020

PUBLISHED 11 September 2020

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.

PAPER

Machine learning in physics: the pitfalls of poisoned training sets

Chao Fang[1,3], **Amin Barzeger**[1,2] **and Helmut G Katzgraber**[2]

[1] Department of Physics and Astronomy, Texas A&M University, College Station, Texas TX 77843-4242, United States of America
[2] Microsoft Quantum, Microsoft, Redmond, Washington WA 98052, United States of America
[3] Author to whom any correspondence should be addressed.

E-mail: fangchao118@gmail.com

Keywords: poisoned training sets, spin glasses, phase transitions, condensed matter physics

Abstract

Known for their ability to identify hidden patterns in data, artificial neural networks are among the most powerful machine learning tools. Most notably, neural networks have played a central role in identifying states of matter and phase transitions across condensed matter physics. To date, most studies have focused on systems where different phases of matter and their phase transitions are known, and thus the performance of neural networks is well controlled. While neural networks present an exciting new tool to detect new phases of matter, here we demonstrate that when the training sets are poisoned (i.e. poor training data or mislabeled data) it is easy for neural networks to make misleading predictions.

1. Introduction

Machine learning methods [1–3] have found applications in condensed matter physics detecting phases of matter and transitions between these on both quantum and classical systems (see, for example, references [4–9]). Different approaches exist, such as lasso [10, 11], sparse regression [12, 13], classification and regression trees [14–16], as well as restricted Boltzmann machines [51], boosting and support vector machines [17–21]. Neural networks [22, 23] are the most versatile and powerful tools, which is why they are commonly used in scientific applications.

Convolutional neural networks (CNNs), in particular, are specialized neural networks for processing data with a grid-like topology. Familiar examples include time-series data, where samples are taken in intervals, and images (two-dimensional data sets). The primary difference between neural networks and convolutional neural networks lies in how hidden layers are managed. In CNNs, a *convolution* is applied to divide the feature space into smaller sections emphasizing local trends. Because of this, CNNs are ideally-suited to study physical models on hypercubic lattices. Recently, it was demonstrated that CNNs can be applied to the detection of phase transitions in Edwards-Anderson Ising spin glasses on cubic lattices [24]. It was shown that the critical behavior of a spin glass with bimodal disorder can be inferred by training the model using data that has Gaussian interactions between the spins. The use of CNNs also results in a reduced numerical effort, which means one could potentially access larger system sizes often needed to overcome corrections to scaling in numerical studies. As such, pairing specialized hardware to simulate Ising systems [25–27] with machine learning techniques might one day elucidate properties of spin glasses and related systems. However, as we show in this work, the use of poor input data can result in erroneous or even unphysical results. This (here inadvertent) *poisoning* of the training set is well known in computer science where small amounts of bad data can strongly affect the accuracy of neural network systems. For example, Steinhardt *et al* [28] demonstrated that already small amounts of bad data can result in a sizable drop in the classification accuracy. References [29–31] furthermore demonstrate that data poisoning can have a strong effect in machine learning. Reference [32] focuses on adversarial manipulations [33, 34] of simulational and experimental data in condensed matter physics applications. In particular, they show that changing individual variables (e.g. a pixel in a data set) can generate misleading predictions. This suggests that results from machine learning algorithms sensitively rely on the quality of the training input.

© 2020 The Author(s). Published by IOP Publishing Ltd

Table 1. Parameters for the training samples with Gaussian disorder. L is the linear size of a system with $N = L^3$ spins, N_{sa} is the number of samples, N_{sw} is the number of Monte Carlo sweeps for each of the replicas for a single sample, T_{min} and T_{max} are the lowest and highest temperatures simulated, N_T is the number of temperatures used in the parallel tempering Monte Carlo method for each system size L, and N_{con} is the number of configurational overlaps for a given temperature in each instance.

L	N_{sa}	N_{sw}	T_{min}	T_{max}	N_T	N_{con}
8	20000	50000	0.80	1.21	20	100
10	10000	40000	0.80	1.21	20	100
12	20000	655360	0.80	1.21	20	100
14	10000	1050000	0.80	1.21	20	100
16	5000	1050000	0.80	1.21	20	100

In this work, we demonstrate that the use of poorly-thermalized Monte Carlo data or simply mislabeled data can result in erroneous estimates of the critical temperatures of Ising spin-glass systems. As such, we focus less on adversarial cases, but more on accidental cases of poor data preparation. We train a CNN with data from a Gaussian Ising spin glass in three space dimensions and then use data generated for a bimodal Ising spin glass to predict the transition temperature of the same model system, albeit with different disorder. In addition, going beyond the work presented in reference [32], we introduce an analysis pipeline that allows for the precise determination of the critical temperature. While good data results in a relatively accurate prediction, the use of poorly-thermalized or mislabeled data produce misleading results. This should serve as a cautionary tale when using machine learning techniques for physics applications.

The paper is structured as follows. In section 2 we introduce the model used in the study, as well as simulation parameters for both training and prediction data. In addition, we outline the implementation of the CNN as well as the approach used to extract the thermodynamic critical temperature, followed by results and concluding remarks.

2. Model and numerical details

To illustrate the effects of poisoned training sets we study the three-dimensional Edwards-Anderson Ising spin glass [35–39] with a neural network implemented in TensorFlow [40]. The model is described by the Hamiltonian

$$\mathcal{H} = -\sum_{\langle i,j \rangle} J_{ij} s_i s_j, \tag{1}$$

where each J_{ij} is a random variable drawn from a given symmetric probability distribution, either bimodal, i.e. ± 1 with equal probability, or Gaussian with zero mean and unit variance. In addition, $s_i = \pm 1$ represent Ising spins, and the sum is over nearest neighbors on a cubic lattice with N sites.

Because spin glasses do not exhibit spatial order below the spin-glass transition, we measure the site-dependent spin overlap [41–43]

$$q_i = S_i^{\alpha} S_i^{\beta}, \tag{2}$$

between replicas α and β. In the overlap space, the system is reminiscent of an Ising ferromagnet, i.e. approaches for ferromagnetic systems introduced in references [6, 7] can be used. For low temperatures, $q = (1/N)\sum_i q_i \to 1$, whereas for $T \to \infty$, $q \to 0$. For an infinite system, q abruptly drops to zero at the critical temperature T_c. Therefore, the overlap space is well suited to detect the existence of a phase transition in a disordered system, even beyond spin glasses. In the overlap space, the spin-glass phase transition can be visually seen as the formation of disjoint islands with identical spin configurations. As such, the problem of phase identification in physical systems is reminiscent of an image classification problem where CNN's are shown to be highly efficient compared to fully-connected neural networks (FCN).

2.1. Data generation

We use parallel tempering Monte Carlo [44] to generate configurational overlaps. Details about the parameters used in the Monte Carlo simulations are listed in table 1 for the training data with Gaussian disorder. The parameters for the prediction data with bimodal disorder are listed in table 2.

2.2. CNN implementation

We use the same amount of instances used in references [45] with 100 configurational overlaps at each temperature for each instance. Because the transition temperature with Gaussian disorder is $T_c \approx 0.95$

Table 2. Parameter for the prediction samples with bimodal disorder. L is the linear size of the system, N_{sa} is the number of samples, N_{sw} is the number of Monte Carlo sweeps for each of the replicas of a single sample, T_{min} and T_{max} are the lowest and highest temperatures simulated, N_T is the temperature numbers used in parallel tempering method for each linear system size L, and N_{con} is the number of configurational overlaps for a given temperature in each instance.

L	N_{sa}	N_{sw}	T_{min}	T_{max}	N_T	N_{con}
8	15000	80000	1.05	1.25	12	500
10	10000	300000	1.05	1.25	12	500
12	4000	300000	1.05	1.25	12	500
14	4000	1280000	1.05	1.25	12	500
16	4000	1280000	1.05	1.25	12	500

Table 3. CNN architecture, parameters, and hardware details.

Number of Layers	1
Channels in each layer	5
Filter size	$3 \times 3 \times 3$
stride	2
Activation function	ReLU
Optimizer	AdamOptimizer(10^{-4})
Batch size	10^3
Iteration	10^4
Software	TensorFlow (Python)
Hardware	Lenovo x86 HPC cluster with a dual-GPU NVIDIA Tesla K80 GPU and 128 GB RAM

[45–47], following references [6–8] for the training data, we label the convolutional overlaps with temperatures above 0.95 as '1' and those from temperatures below 0.95 as '0'.

The parameters for the architecture of the convolutional neural network are listed in table 3. We inherit the structure with a single layer from reference [8]. All the parameters are determined by extra validation sample sets, which are also generated from Monte Carlo simulations.

Note that we use between 4000 and 10000 disorder instances for the bimodal prediction data, which is approximately 1/3 of the numerical effort needed when estimating the phase transition directly via a finite-scaling analysis of Monte Carlo data, as done for example in reference [45]. As such, *pairing high-quality Monte Carlo simulations with machine learning techniques can result in large computational cost savings.*

2.3. Data analysis

Because the configurational overlaps (equation (2)) include the information about phases, we expect that different phases have different overlap patterns similar to grid-like graphs. Therefore, in the region of a specific phase, it is reasonable to expect that the classification probability for the CNN to identify the phase correctly should be larger than 50%. As such, it can be expected that when the classification probability is 0.5, the system is at the system-size-dependent critical temperature. A thermodynamic estimate can then obtained via the finite-size scaling method presented below.

Let us define the classification probability as a function of temperature and system size: $p(T, L)$ which can be used as a dimensionless quantity to describe the critical behavior. From the scaling hypothesis, we expect $p(T, L)$ to have the following behavior in the vicinity of the critical temperature T_c:

$$\langle p(T,L)\rangle = \tilde{F}\left[L^{1/\nu_{ml}}\left(T - T_c\right)\right], \tag{3}$$

where the average is over disorder realizations. Note that the critical exponent ν_{ml} is different from the one calculated using physical quantities. Due to the limited system sizes that we have studied, finite-size scaling must be used to reliably calculate the critical parameters at the thermodynamic limit. Assuming that we are close enough to the critical temperature T_c, the scaling function $\tilde{F}$ in equation (3) can be expanded to a third-order polynomial in $x = L^{1/\nu_{ml}}\left(T - T_c\right)$.

$$\langle p(T,L)\rangle \sim p_0 + p_1 x + p_2 x^2 + p_3 x^3. \tag{4}$$

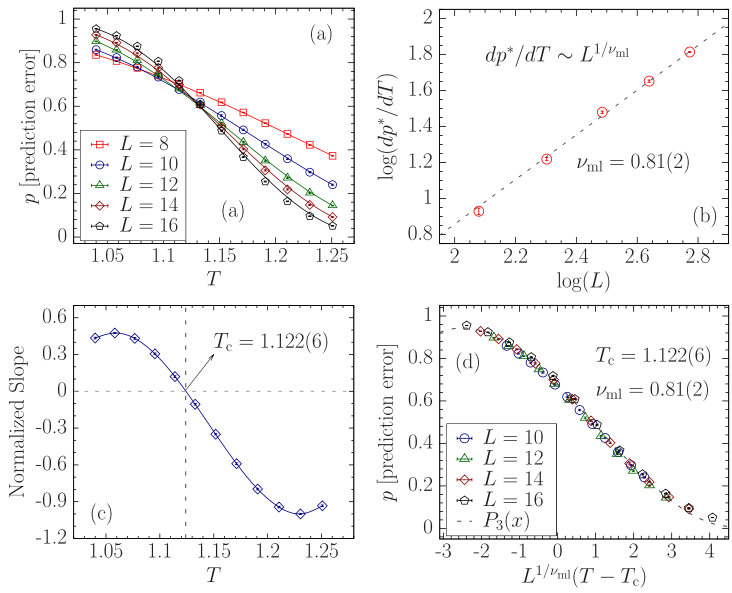

Figure 1. Classification probabilities for different linear system sizes L as a function of temperature T for the prediction of the critical temperature of the bimodal Ising spin glass via a CNN trained with data from a Gaussian distribution. (a) Prediction probability for different system sizes L near the phase transition temperature. The different data sets cross at $T_c \sim 1.122$. (b) Measurement of ν_{ml} by performing a linear fit in a double-logarithmic scale using the extremum points of the derivative of the prediction error with respect to the temperature. (c) Estimate of the critical temperature T_c using the coefficient of the linear term in equation (4) (normalized to 1) with $L^{1/\nu_{ml}}$ as the independent variable. The vertical dashed line shows the temperature where the slope vanishes, which corresponds to T_c. (d) Finite-size scaling of the data using the previously-estimated value of ν_{ml} and T_c. The data collapse onto a universal curve indicating that the estimates are accurate.

First, we evaluate ν_{ml} by noting that to the leading order in x, the derivative of $\langle p(T,L)\rangle$ in equation (4) with respect to temperature has the following form:

$$\frac{d\langle p(T,L)\rangle}{dT} \sim L^{1/\nu_{ml}}\left[p_1 + 2p_2 L^{1/\nu_{ml}}(T-T_c) + 3p_3 L^{2/\nu_{ml}}(T-T_c)^2\right]. \tag{5}$$

Therefore, the extremum point of $\frac{d\langle p(T,L)\rangle}{dT}$ scales as

$$\frac{d\langle p(T,L)\rangle}{dT}\Big|_{T=T^*} \sim L^{1/\nu_{ml}}. \tag{6}$$

A linear fit in a double-logarithmic scale then produces the value of ν_{ml} (slope of the straight line), which is subsequently used to estimate T_c. To do so, we turn back to equation (4) where we realize that the coefficient of the linear term in $L^{1/\nu_{ml}}$ as the independent variable is proportional to $(T-T_c)$ that changes sign at $T=T_c$. Alternatively, we can vary T_c until the data for all system sizes collapse onto a common third-order polynomial curve. This is true because the scaling function $\tilde{F}$ as a function of $L^{1/\nu_{ml}}(T-T_c)$ is universal. The error bars can be computed using the bootstrap method.

3. Results using data without poisoning

Figure 1 shows results from the CNN trained with well-prepared (thermalized) data from a Gaussian distribution, predicting the phase transition of data from a Bimodal disorder distribution. Figure 1(a) shows the prediction probabilities for different linear system sizes L as a function of temperature T. The curves cross the $p=0.5$ line in the region of the transition temperature for the bimodal Ising spin glass. Figures 1(b) and (c) show the estimates of the exponent ν_{ml} and the critical temperature T_c, respectively, using the methods developed in section 2.3. The critical temperature $T_c = 1.122(6)$ is in good agreement with previous estimates (see, for example, reference [45]). Finally, in figure 1(d), the data points are plotted as a function of the reduced variable $x = L^{1/\nu_{ml}}(T-T_c)$ using the estimated values of the critical parameters. The universality of the scaling curve underlines the accuracy of the estimates.

4. Results using poisoned training sets

Although we have shown that the prediction from a convolutional neural network can be precise, we still need to test how poisoned data sets impact the final prediction. First, we randomly mix the classification

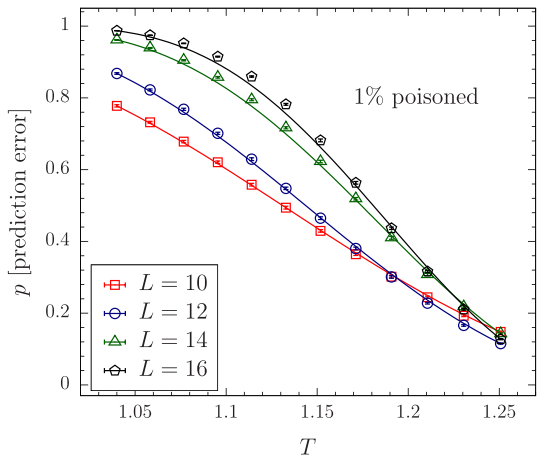

Figure 2. Classification probabilities for different system sizes L for an Ising spin glass with bimodal disorder. 1% of the labels have been mixed on average. There is no clear sign of the transition.

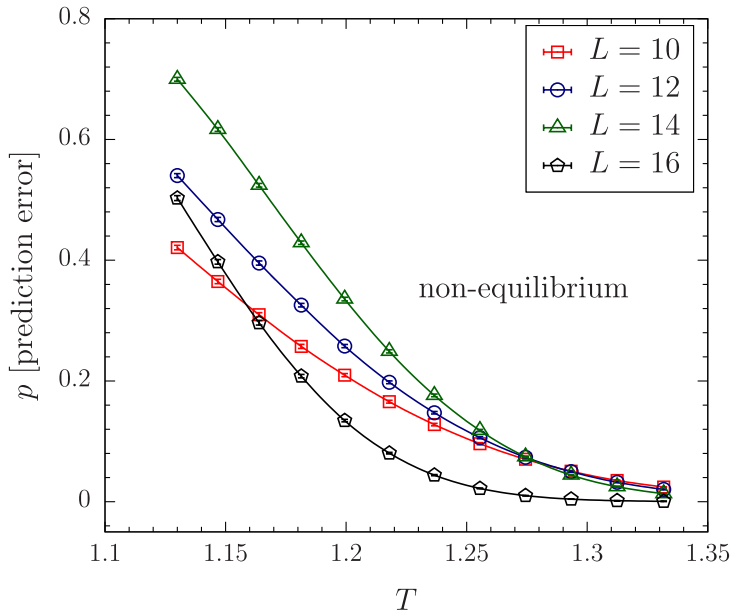

Figure 3. Classification probabilities for different system sizes L for an Ising spin glass with bimodal disorder. The Gaussian training data are not thermalized. There is no clear sign of a phase transition.

labels of the training sample with a probability of 1%, i.e. with a training set of 100 samples, this means only one mislabeled sample on average. Then we train the network and use the same samples in the prediction stage. Compared to figure 1, figure 2 shows no clear sign of a phase transition. This means that mislabeling a very small portion of the training data can strongly affect the outcome. Given the hierarchical structure of CNNs, errors can easily be amplified in propagation [48, 49], which is a possible explanation of the observed behavior.

Finally, we test the effects of poorly prepared training data–in this case, the training data are not properly thermalized. Figure 3 shows the prediction results using data with only 50% of the Monte Carlo sweeps needed for thermalization of the Gaussian training samples. Although 50% might seem extreme at first sight, it is important to emphasize that thermalization times (as well as time-to-solution) are typically distributed according to fat-tail distributions [50]. In general, users perform at least a factor 2 of additional thermalization to ensure most instances are in thermal equilibrium. As in the case where the labels were mixed, a transition cannot be clearly identified. This is strong indication that the training data need to be carefully prepared.

We have also studied the effects of poorly-thermalized prediction data paired with well-thermalized training data (not shown). In this case, the impacts on the prediction probabilities are small but not negligible.

5. Discussion

We have studied the effects of poisoned data sets when training CNNs to detect phase transitions in physical systems. Our results show that good training sets are a necessary requirement for good predictions. Small perturbations in the training set can lead to misleading results.

We do note, however, that we might not have selected the best parameters for the CNN. Using cross-validation or bootstrapping might allow for a better tuning of the parameters and thus improve the quality of the predictions. Furthermore, due to the large number of predictors, overfitting is possible. This, however, can be alleviated by the introduction of penalty terms. Finally, the use of other activation functions and optimizers can also impact the results. This, together with the sensitivity towards the quality of the training data that we find in this work suggest that machine learning techniques should be used with caution in physics applications. Garbage in, garbage out …

Acknowledgments

We would like to thank Humberto Munoz Bauza and Wenlong Wang for fruitful discussions. This work is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via MIT Lincoln Laboratory Air Force Contract No. FA8721-05-C-0002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation therein. We thank Texas A&M University for access to their Terra cluster.

ORCID iDs

Chao Fang https://orcid.org/0000-0002-6626-6587
Helmut G Katzgraber https://orcid.org/0000-0003-3341-9943

References

[1] Haykin S O 2009 *Neural Networks and Learning Machines* (New York: Pearson)
[2] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press) (www.deeplearningbook.org)
[3] Bishop C 2006 *Pattern Recognition and Machine Learning* (New York: Springer)
[4] Ronhovde P, Chakrabarty S, Hu D, Sahu M, Sahu K K, Kelton K F, Mauro N A and Nussinov Z 2011 *The European Physical Journal E* **34** 105
[5] Nussinov Z, Ronhovde P, Hu D, Chakrabarty S, Sun B, Mauro N A and Sahu K K 2016 *Information Science for Materials Discovery and Design* (Cham: Springer Int. Publishing)ed Lookman T, Alexander F J and Rajan K p 115 Springer Series in Materials Science
[6] Carrasquilla J and Melko R G 2017 *Nat. Phys.* **13** 431
[7] Ch'ng K, Carrasquilla J, Melko R G and Khatami E 2017 *Phys. Rev.* X **7** 031038
[8] Tanaka A and Tomiya A 2017 *J. Phys. Soc. Jpn* **86** 063001
[9] Kashiwa K, Kikuchi Y and Tomiya A 2019 *Prog. Theor. Exp. Phys.* **2019** 083A04
[10] Santosa F and Symes W W 1986 *SIAM J. Sci. Stat. Comput.* **7** 1307
[11] Tibshirani R 1996 *J. Royal Stat. Soc.* B **58** 267–88
[12] Mateos G, Bazerque J A and Giannakis G B 2010 *Trans. Sig. Proc.* **58** 5262
[13] Quinonero Candela J and Rasmussen C 2005 *Journal of Machine Learning Research* **6** 1935–59
[14] Rokach L and Maimon O 2014 *Data Mining With Decision Trees: Theory and Applications* (River Edge, NJ, USA: World Scientific Publishing Co., Inc.) 2nd ed
[15] Shalev-Shwartz S and Ben-David S 2014 *Understanding Machine Learning: From Theory to Algorithms* (Cambridge: Cambridge University Press)
[16] Mehta D and Raghavan V 2002 *Theor. Comput. Sci.* **270** 609
[17] James G, Witten D, Hastie T and Tibshirani R 2013 *An Introduction to Statistical Learning With Application in R* (Berlin: Springer)
[18] Hsu C, Chang C and Lin C 2010 *A practical guide to support vector classification* (Taipei: National Taiwan University)
[19] Platt J C 1999 *Fast Training of Support Vector Machines Using Sequential Minimal Optimization* (Cambridge, MA: MIT Press) p 185
[20] Widodo A and Yang B-S 2007 *Mech. Syst. Signal Process.* **21** 2560
[21] Joachims T 1998 *Proc. of the 10th Conf. on Machine Learning* (Berlin: Springer) p 137
[22] LeCun Y and Bengio Y 1998 *Convolutional Networks for Images, Speech and Time Series* (Cambridge, MA: MIT Press) p 255
[23] Zhang W, Itoh K, Tanida J and Ichioka Y 1990 *Appl. Opt.* **29** 4790
[24] Munoz-Bauza H, Hamze F and Katzgraber H G 2020 *J. Stat. Mech.* **2020** 073302
[25] Alvarez Baños R *et al* 2010 *J. Stat. Mech.* **2010** P06026
[26] Baños R A *et al* 2012 *Proc. Natl. Acad. Sci. U.S.A.* **109** 6452
[27] Baity-Jesi M *et al* 2014 *Phys. Rev.* E **89** 032140
[28] Steinhardt J, Koh P W and Liang P 2017 arXiv:cs/1706.03691
[29] Jagielski M, Oprea A, Biggio B, Liu C and Nita-Rotaru B and Li C 2018 arXiv:cs/1804.00308
[30] Alfeld R, Zhu X and Barford P 2016 *AAAI* 1452
[31] Shi Y, Erpek T, Sagduyu Y E and Li J H 2019 arXiv:cs/1901.09247

[32] Jiang S, Lu S and Deng D-L 2019 Vulnerability of machine learning phases of matter arXiv:cond-mat/1910.13453
[33] Nelson B, Barreno F, Chi F J, Joseph A D, Rubinstein B I, Saini U, Sutton C, Tygar J and Xia K 2008 *Proc. First USENIX Workshop on Large-Scale Exploits and Emergent Threats, LEET*
[34] Newell A, Potharaju L, Xiang L and Nita-Rotaru C 2014 *Proc. Workshop on Artificial Intelligence and Security AISec 2014*
[35] Edwards S F and Anderson P W 1975 *J. Phys. F.: Met. Phys.* **5** 965
[36] Binder K and Young A P 1986 *Rev. Mod. Phys.* **58** 801–976
[37] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[38] Young A P ed 1998 *Spin Glasses and Random Fields* (Singapore: World Scientific)
[39] Stein D L and Newman C M 2013 *Spin Glasses and Complexity, Primers in Complex Systems* (Princeton, NJ: Princeton University Press)
[40] Abadi M *et al* 2016 *TensorFlow: A System for Large-Scale Machine Learning* (http://tensorflow.org)
[41] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[42] Parisi G 1980 *J. Phys.* A **13** 1101
[43] Parisi G 1983 *Phys. Rev. Lett.* **50** 1946
[44] Hukushima K and Nemoto K 1996 *J. Phys. Soc. Jpn.* **65** 1604
[45] Katzgraber H G, Körner M and Young A P 2006 *Phys. Rev.* B **73** 224432
[46] Marinari E, Parisi G and Ruiz-Lorenzo J J 1998 *Phys. Rev.* B **58** 14852
[47] Katzgraber H G and Campbell I A 2005 *Phys. Rev.* B **72** 014462
[48] Rumelhart D E, Hinton G E and Williams R J 1988 *Learning Representations by Back-Propagating Errors* (Cambridge, MA: MIT Press) p 696
[49] Hecht-Nielsen R 1992 *Theory of the Backpropagation Neural Network* (Orlando, FL, USA: Harcourt Brace & Co.) p 65
[50] Steiger D S, Rønnow T F and Troyer M 2015 *Phys. Rev. Lett.* **115** 230501
[51] Azizi A and Pleimling M 2020 Machine learning generated configurations in presence of a conserved quantity: a cautionary tale arXiv:2007.09764