

Title: Quantum HyperNetworks: Training Binary Neural Networks in Quantum Superposition

Speakers: Estelle Inack

Collection: Machine Learning for Quantum Many-Body Systems

Date: June 16, 2023 - 11:15 AM

URL: <https://pirsa.org/23060035>

**Abstract:** Binary neural networks, i.e., neural networks whose parameters and activations are constrained to only two possible values, offer a compelling avenue for the deployment of deep learning models on energy- and memory-limited devices. However, their training, architectural design, and hyperparameter tuning remain challenging as these involve multiple computationally expensive combinatorial optimization problems. Here we introduce quantum hypernetworks as a mechanism to train binary neural networks on quantum computers, which unify the search over parameters, hyperparameters, and architectures in a single optimization loop. Through classical simulations, we demonstrate that our approach effectively finds optimal parameters, hyperparameters and architectural choices with high probability on classification problems including a two-dimensional Gaussian dataset and a scaled-down version of the MNIST handwritten digits. We represent our quantum hypernetworks as variational quantum circuits, and find that an optimal circuit depth maximizes the probability of finding performant binary neural networks. Our unified approach provides an immense scope for other applications in the field of machine learning.

# Quantum HyperNetworks: Training Binary Neural Networks in Quantum Superposition

**Estelle Inack**

*Machine Learning for Quantum Many-Body Systems*

16<sup>th</sup> June 2023

*J. Carrasquilla, M. Hibat-Allah, E.M.J, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1*



# ChatGPT: The Latest AI Sensation

Business

## Can the new AI tool ChatGPT replace human work? Judge for yourself



New artificial intelligence tool can respond to a human prompt **The technology behind it is getting even more powerful**

By Samantha Murphy Kelly, CNN Business  
Updated 4:42 AM EDT, Wed March 15, 2023

ChatGPT  
moment

## TECH What Is ChatGPT? Unprecedented consumer interest About the AI Chatbot

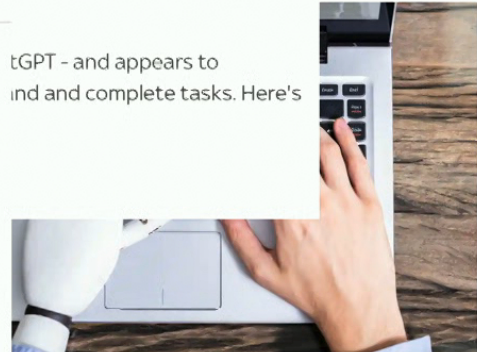
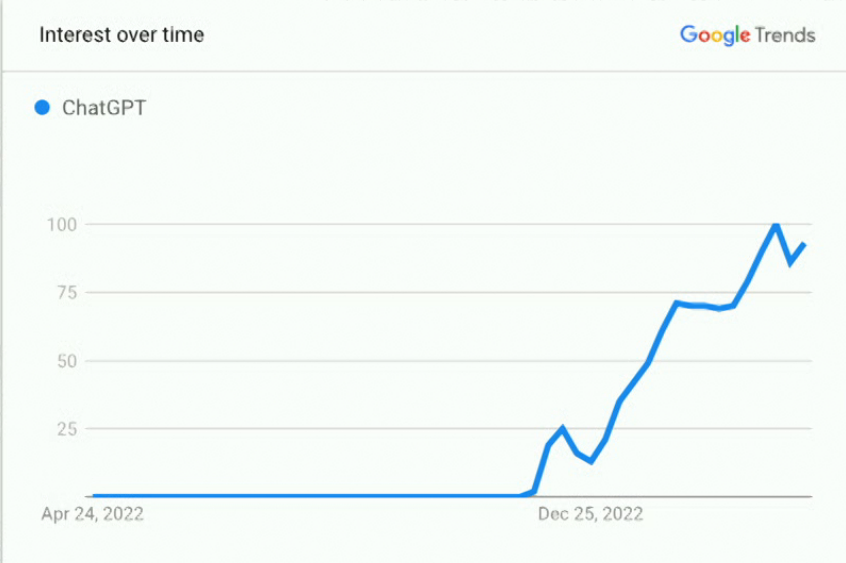
OpenAI's chatbot and Microsoft's conversational Bing have triggered a new AI race that may reshape the future of work

By [Karen Hao](#) [Follow](#)  
Updated April 11, 2023 at 8:44 pm ET

Explainer

## What is AI chatbot phenomenon ChatGPT and could it replace humans?

## What is OpenAI's GPT-4 and how does it



ChatGPT - and appears to understand and complete tasks. Here's

halled as a potential game-changer in the world of AI. Popov/Alamy

## The Brilliance and Weirdness of ChatGPT

A new chatbot from OpenAI is inspiring awe, fear, stunts and attempts to circumvent its guardrails.

By Kevin Roose





# Some issues of generative AI

CHRIS STOEL-WALKER BUSINESS FEB 18, 2023 7:00 AM

## The Generative AI Race Has a Dirty Secret

Integrating large language models into search engines could mean a fivefold increase in computing power and huge carbon emissions.

---

While neither OpenAI nor Google, have said what the computing cost of their products is, [third-party analysis](#) by researchers estimates that the training of GPT-3, which ChatGPT is partly based on, consumed 1,287 MWh, and led to emissions of more than 550 tons of carbon dioxide equivalent—the same amount as a single person taking 550 roundtrips between New York and San Francisco.





# Some issues of generative AI

CHRIS STEIGEL-WALKER BUSINESS FEB 18, 2023 7:00 AM

## The Generative AI Race Has a Dirty Secret

Integrating large language models into search engines could mean a fivefold increase in computing power and huge carbon emissions.

Consumption	CO <sub>2</sub> e (lbs)	
Air travel, 1 passenger, NY↔SF	1984	
Human life, avg, 1 year	11,023	
American life, avg, 1 year	36,156	
Car, avg incl. fuel, 1 lifetime	126,000	
<hr/>		
<b>Training one model (GPU)</b>		
NLP pipeline (parsing, SRL)	39	t of their products
w/ tuning & experimentation	78,468	PT-3, which
Transformer (big)	192	ons of more than
w/ neural architecture search	626,155	person taking 550

Table 1: Estimated CO<sub>2</sub> emissions from training common NLP models, compared to familiar consumption.<sup>1</sup>

[1] E. Strubell, A. Ganesh and A. McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, arXiv:1906.02243

Computer Science > Machine Learning

[Submitted on 6 Apr 2023]

## Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

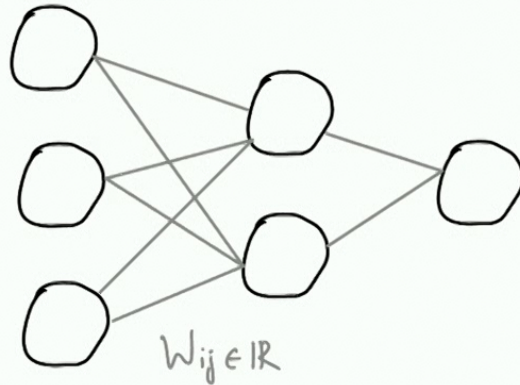
Pengfei Li, Jianyi Yang, Mohammad A. Islam, Shaolei Ren

The growing carbon footprint of artificial intelligence (AI) models, especially large ones such as GPT-3 and GPT-4, has been undergoing public scrutiny. Unfortunately, however, the equally important and enormous water footprint of AI models has remained under the radar. For example, training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly consume 700,000 liters of clean freshwater (enough for producing 370 BMW cars or 320 Tesla electric vehicles) and the water consumption would have been tripled if training were done in Microsoft's Asian data centers, but such information has been kept as a secret. This is extremely concerning, as freshwater scarcity has become one of the most pressing challenges shared by all of us in the wake of the rapidly growing population, depleting water resources, and aging water infrastructures. To respond to the global water challenges, AI models can, and also should, take social responsibility and lead by example by addressing their own water footprint. In this paper, we provide a principled methodology to estimate fine-grained water footprint of AI models, and also discuss the unique spatial-temporal diversities of AI models' runtime water efficiency. Finally, we highlight the necessity of holistically addressing water footprint along with carbon footprint to enable truly sustainable AI.

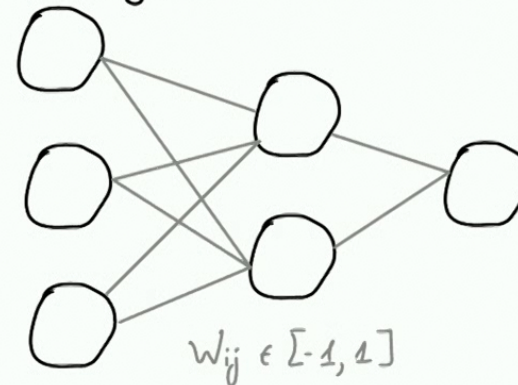


# Binary Neural Networks (BiNNs)

Standard NN



Binary NN

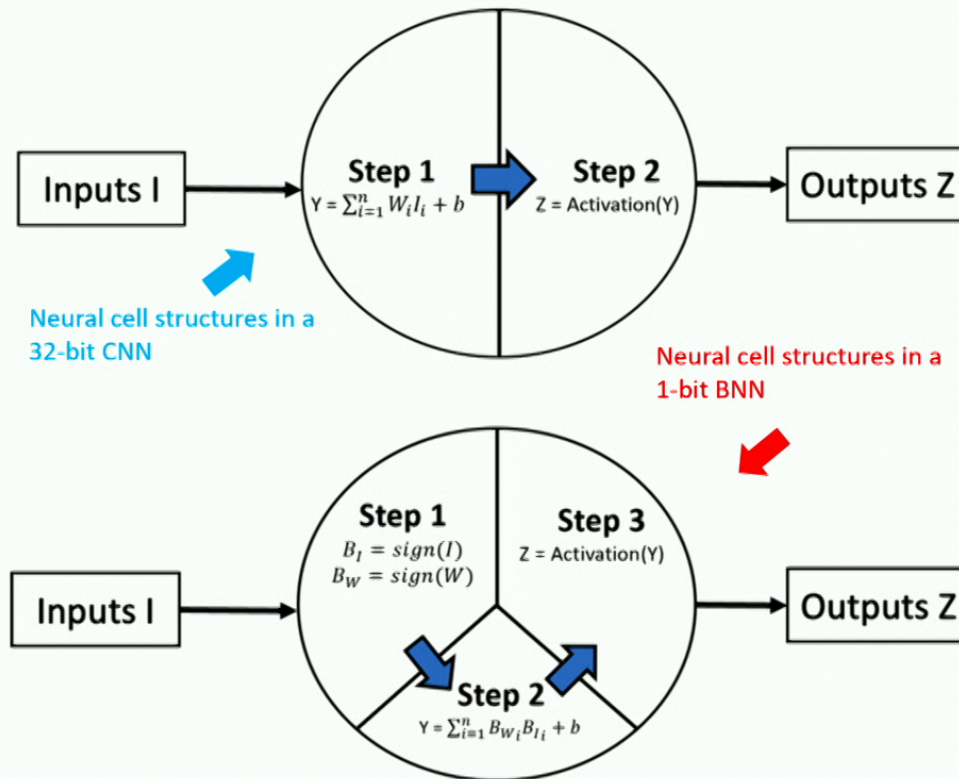


- BiNNs are neural networks with weights and activations constrained to two possible values (e.g. -1 and 1.)[1]
- Training is performed with real weights but binary activation functions – inference is made on binary weights

[1] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, *Binarized neural networks*, Advances in neural information processing systems 29 (2016)



# BiNNs — Example of Matrix-Vector operation

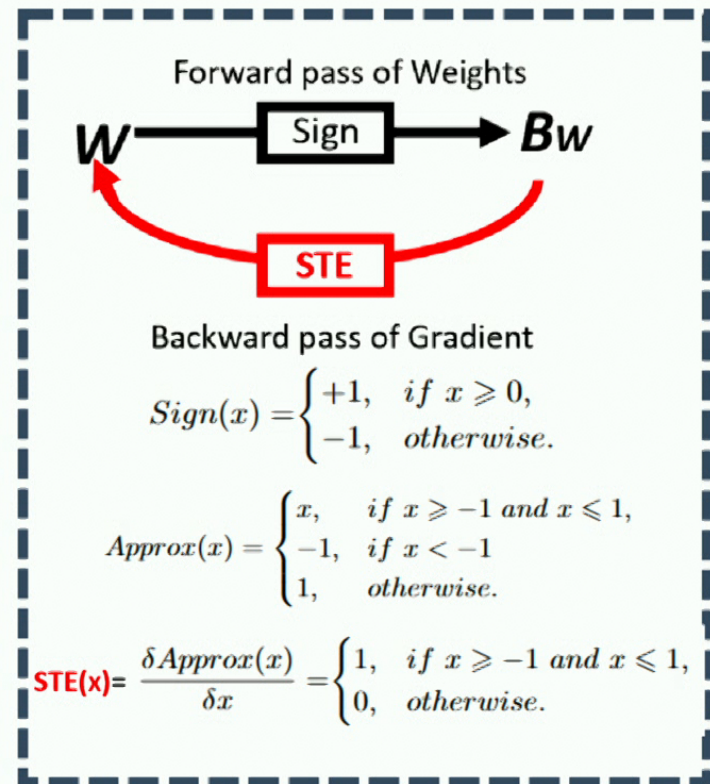
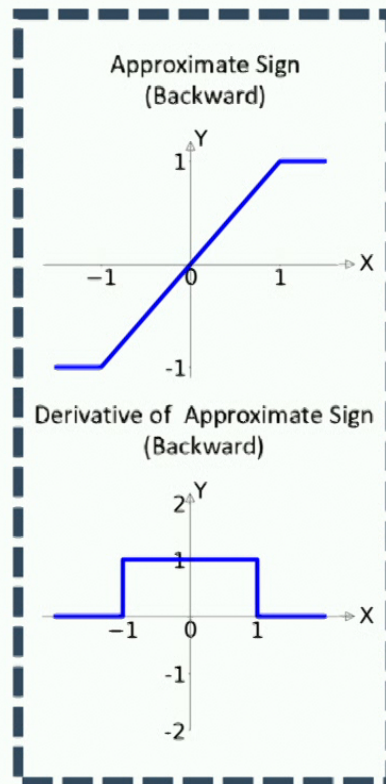
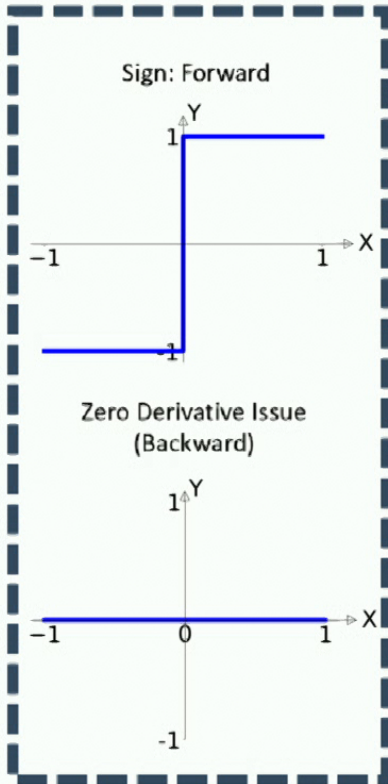


- Picture taken from C. Yuan and S. S. Aghaian, arXiv:2110.06804v4
- Inference step can leverage bitwise operations like XNOR and POPCOUNT





# BiNNs – Training procedure



Picture taken from C. Yuan and S. S. Agaian, arXiv:2110.06804v4

## BiNNs – Some results

**Table 12.** BNN Accuracy comparisons on CIFAR-10

Full Precision Base CNN			
CNN Name		Acc(%)	
VGG-Small(2018)		93.8	
VGG-11(2019)		83.8	
NIN(2019)		84.2	
ResNet-18(2020b)		93.0	
ResNet-20(2020b)		91.7	
WRN-22(2016)		92.62	
WRN-22(4 x Kernel Stage) <sup>1</sup> (2016)		95.75	
BNN Accuracy Perform		Bi-Real-Net(2018)	ResNet-18** <sup>2</sup>
BNN Name	Topology	HadaNet(2019)	89.12(2021)
BNN(2016)	VGG-Small	Customized( $\beta_w=4; \beta_a=4$ )	88.64
XNOR-Net(2016)	VGG-Small	NIN( $\beta_w=4; \beta_a=4$ )	87.33
	WRN-22	Customized( $\beta_w=16; \beta_a=2$ )	89.02
	WRN-22(4 x Kernel Stag	NIN( $\beta_w=16; \beta_a=2$ )	88.74
	ResNet-18	PCNN(2019)	WRN-22
			89.17(J=1) <sup>3</sup>
			WRN-22
			91.27(J=2) <sup>3</sup>
			WRN-22
			92.79(J=4) <sup>3</sup>
			WRN-22(4 x Kernel Stage) <sup>1</sup>
			94.31(J=1) <sup>3</sup>
			WRN-22(4 x Kernel Stage) <sup>1</sup>
			95.39(J=4) <sup>3</sup>

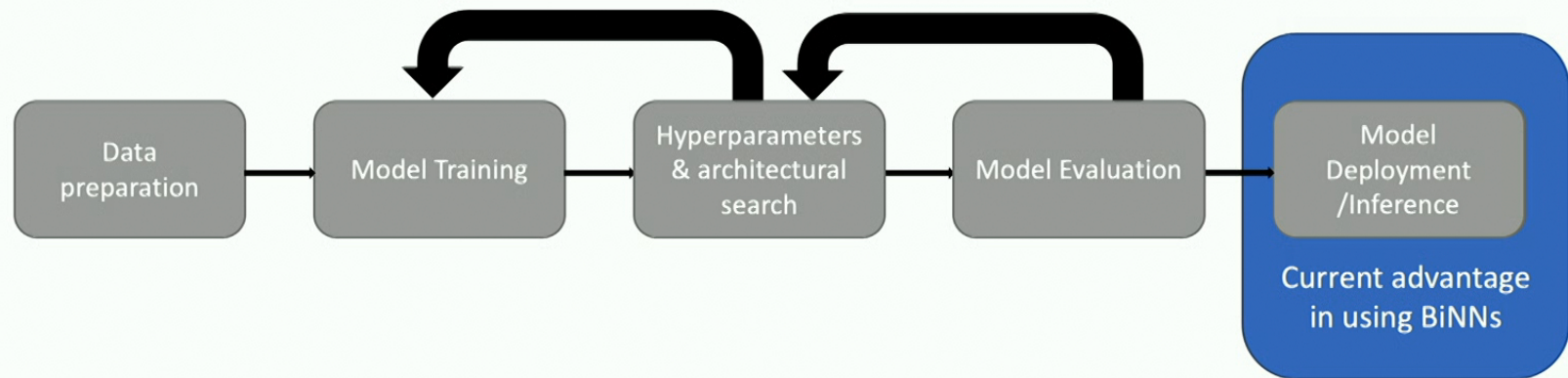
- They have demonstrated more than 100x energy efficiency on FPGAs compared to GPUs for comparable accuracy[1].
- And 32x memory with 2x computational speed compared to AlexNet[2]
- Improvements still needed to reach full-precision network accuracy.

Picture taken from Yuan and S. S. Aghajani, arXiv:2110.06804v4

[1] Gao, J.; Liu, Q.; Lai, J. An Approach of Binary Neural Network Energy-Efficient Implementation. Electronics 2021, 10, 1830.  
 [2] M. Rastegari, V. Ordonez, J. Redmon, A. Farhadi, XNOR-Net: ImageNet Classification Using Binary CNNs, arXiv:1603.05279v4



## Typical Machine Learning Pipeline



- **Can we provide BiNN advantage before the inference step?**
- Nested training with hyperparameter and architectural search causes three orders of magnitude more costs[1]  
**Can we reduce BiNN training to a single optimization loop?**

[1] E. Strubell, A. Ganesh and A. McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, arXiv:1906.02243






# Quantum computers energy advantage

## PAPER

### Establishing the quantum supremacy frontier with a 281 Pflop/s simulation

Benjamin Villalonga<sup>1,2,3</sup>, Dmitry Lyakh<sup>4,5</sup>, Sergio Boixo<sup>6</sup>, Hartmut Neven<sup>6</sup>, Travis S Humble<sup>4</sup>, Rupak Biswas<sup>1</sup>, Eleanor G Rieffel<sup>1</sup>, Alan Ho<sup>6</sup> and Salvatore Mandrà<sup>1,7,8</sup> 

#### Abstract

Noisy intermediate-scale quantum (NISQ) computers are entering an era in which they can perform computational tasks beyond the capabilities of the most powerful classical computers, thereby achieving 'quantum supremacy', a major milestone in quantum computing. NISQ supremacy requires comparison with a state-of-the-art classical simulator. We report HPC simulations of hard random quantum circuits (RQC), which have been recently used as a benchmark for the first experimental demonstration of quantum supremacy, sustaining an average performance of 281 Pflop/s (true single precision) on Summit, currently the fastest supercomputer in the world. These simulations were carried out using qFlex, a tensor-network-based classical high-performance simulator of RQCs. **Our results show an advantage of many orders of magnitude in energy consumption of NISQ devices over classical supercomputers.** In addition, we propose a standard benchmark for NISQ computers based on qFlex.



# Training BiNNs with quantum computers — previous works

---

## Quantum Annealing Formulation for Binary Neural Networks

---

**Michele Sasdelli**   **Tat-Jun Chin**  
School of Computer Science, The University of Adelaide  
Adelaide SA 5005, Australia  
{michele.sasdelli,tat-jun.chin}@adelaide.edu.au

## Quantum advantage in training binary neural networks

Yidong Liao,<sup>1</sup> Daniel Ebler,<sup>1,2</sup> Feiyang Liu,<sup>1</sup> and Oscar Dahlsten<sup>1,3,4,2,\*</sup>  
<sup>1</sup>*Institute for Quantum Science and Engineering, Department of Physics, Southern University of Science and Technology (SUSTech), Shenzhen, China*  
<sup>2</sup>*Wolfson College, University of Oxford, Linton Road, Oxford OX2 6UD, UK*  
<sup>3</sup>*Center for Quantum Computing, Peng Cheng Laboratory, Shenzhen, 518000, China*  
<sup>4</sup>*London Institute for Mathematical Sciences, 35a South Street Mayfair, London W1K 2XF, UK*

Research Track Paper

KDD '19, August 4–8, 2019, Anchorage, AK, USA

## Training and Meta-Training Binary Neural Networks with Quantum Computing

**Abdulah Fawaz**  
abdulah.fawaz@siemens-healthineers.com  
Siemens Healthineers, Digital Services, Digital Technology and Innovation,  
Princeton, New Jersey, USA

**Paul Klein**  
**Sebastien Plat**  
klein.paul@siemens-healthineers.com  
sebastien.plat@siemens-healthineers.com  
Siemens Healthineers, Digital Services, Digital Technology and Innovation,  
Princeton, New Jersey, USA

**Simone Severini**  
s.severini@ucl.ac.uk  
Department of Computer Science, University College London

**Peter Mountney**  
peter.mountney@siemens-healthineers.com  
Siemens Healthineers, Digital Services, Digital Technology and Innovation,  
Princeton, New Jersey, USA

## QUANTUM-AIDED META-LEARNING FOR BAYESIAN BINARY NEURAL NETWORKS VIA BORN MACHINES

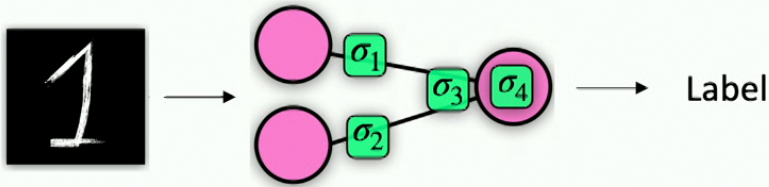
*Ivana Nikoloska and Osvaldo Simeone*

KCLIP, CTR, Department of Engineering, King's College London





# BiNNs in quantum superposition



$\sigma_1$  and  $\sigma_2$  are weights which take value  $\{0,1\}$ .

$\sigma_3$  is a bias

$\sigma_4$  encodes an architectural choice, for e.g. activation function choice

$$f(x; \sigma_4) = \begin{cases} f_1(x) & \text{if } \sigma_4 = 0 \\ f_2(x) & \text{if } \sigma_4 = 1 \end{cases}$$

$$\mathbf{w} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$$

$$C(\mathbf{w}) = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathcal{L}(\text{NN}(\mathbf{x}_i; \mathbf{w}), \mathbf{y}_i)$$

$$\mathbf{w} \rightarrow \hat{\boldsymbol{\sigma}}_z = (\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\sigma}_4)$$

$$|\Psi\rangle = \sum_{\sigma_1, \sigma_2, \sigma_3, \sigma_4} \Psi(\sigma_1, \sigma_2, \sigma_3, \sigma_4) |\sigma_1, \sigma_2, \sigma_3, \sigma_4\rangle$$

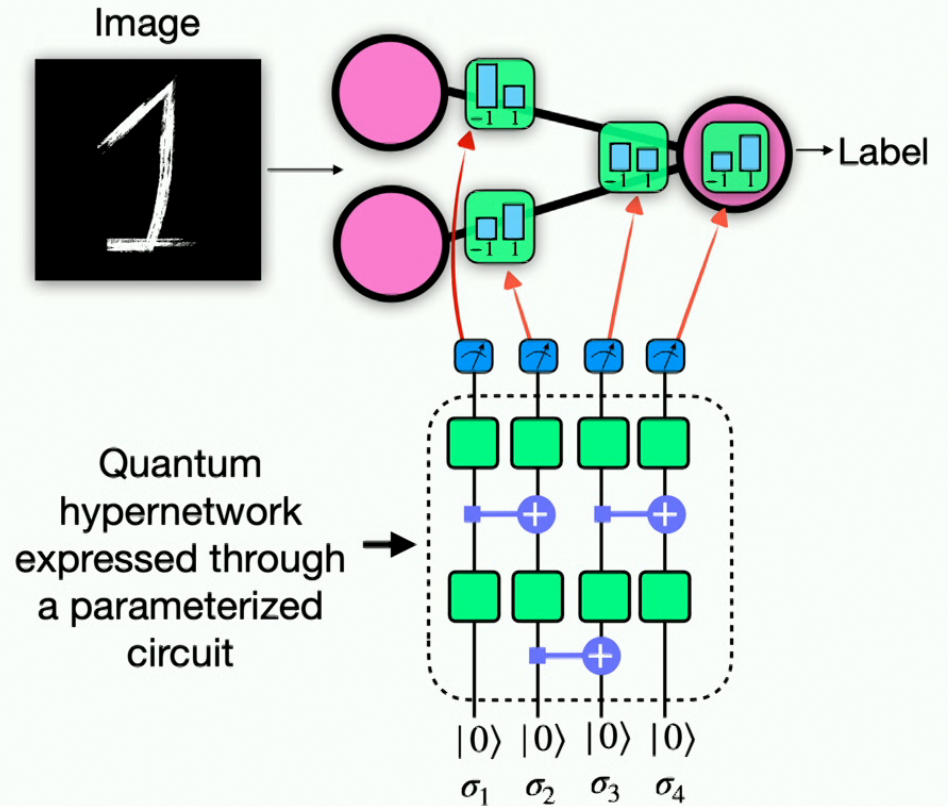
$$|\Psi\rangle = \sum_{\sigma_1, \dots, \sigma_4} \Psi(\sigma_1, \dots, \sigma_4) \left| \begin{array}{c} \text{pink circle} \text{---} \sigma_1 \\ \text{pink circle} \text{---} \sigma_2 \end{array} \text{---} \sigma_3 \text{---} \sigma_4 \right\rangle$$



# Variational Quantum Algorithms

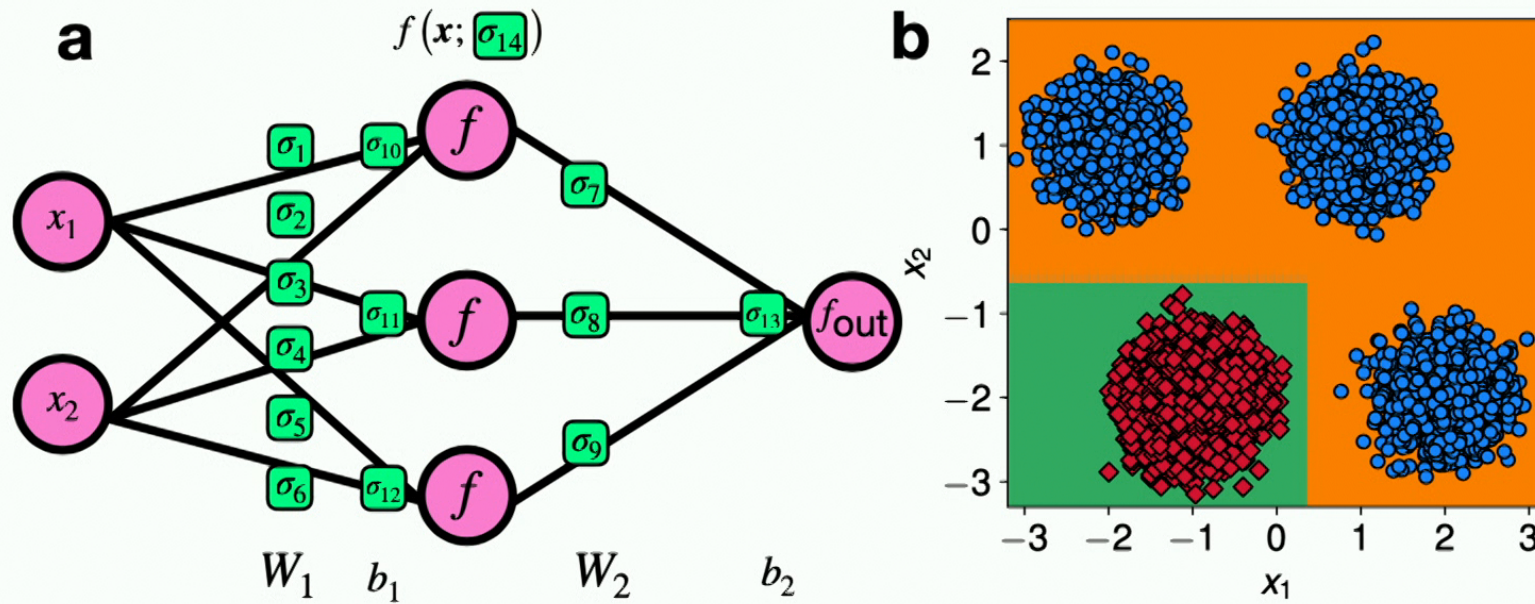
$$\theta^* = \arg \min_{\theta} E(\theta)$$

$$\begin{aligned} E(\theta) &= \langle \Psi_{\theta} | \hat{C} | \Psi_{\theta} \rangle \\ &= \sum_{\sigma_1, \sigma_2, \dots, \sigma_N} |\Psi_{\theta}(\sigma_1, \sigma_2, \dots, \sigma_N)|^2 C(\sigma_1, \sigma_2, \dots, \sigma_N) \\ &= \mathbb{E}_{\sigma \sim |\Psi_{\theta}|^2} [C(\sigma)] \approx \frac{1}{N_{qc}} \sum_{i=1}^{N_{qc}} C(\sigma_i), \end{aligned}$$



J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1

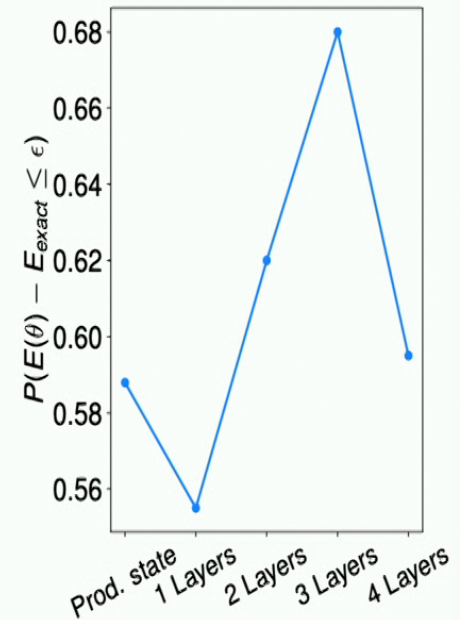
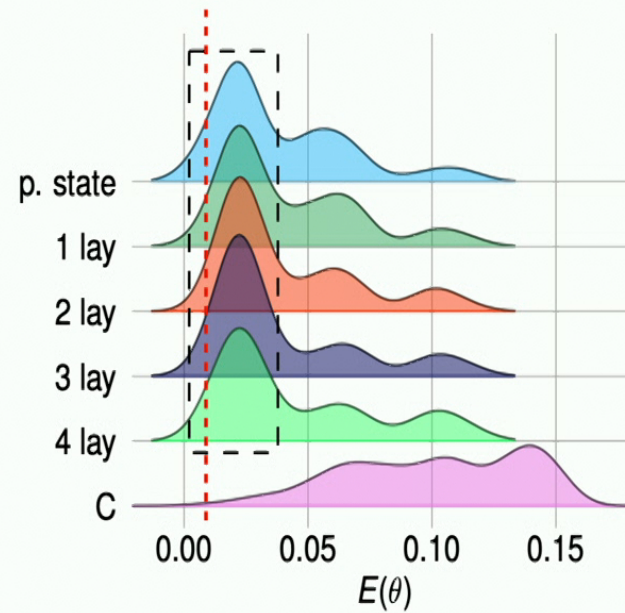
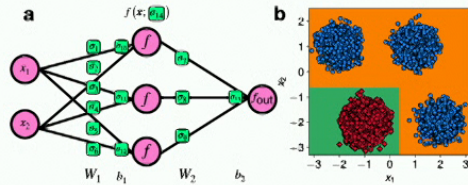
## Results: BiNNs applied to a Gaussian dataset – Full enumeration



J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1



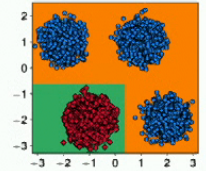
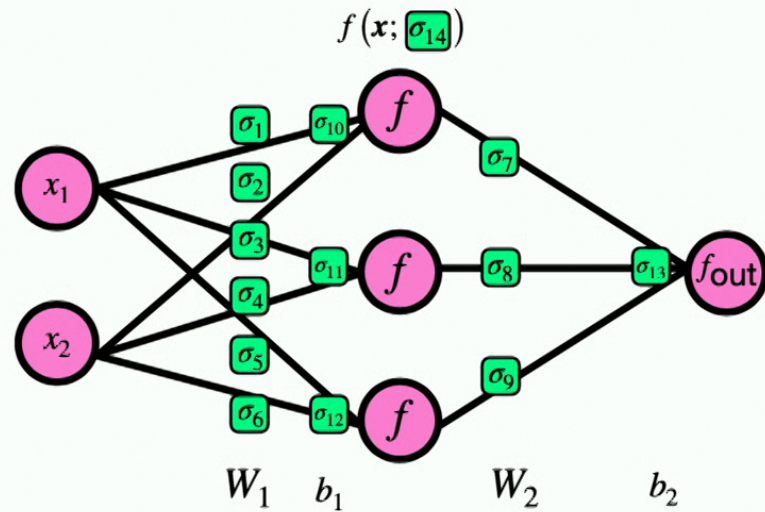
# Results: BiNNs applied to a Gaussian dataset



- **Task:** Binary classification of data points
- **Training objective:** Optimize weights, biases and architectural choice of nonlinearity
- Run optimization at least 200 times and evaluate the probabilities of finding an objective function with value  $E(\theta)$ .
- Compute the probability that  $E(\theta)$  is less than  $\epsilon$ .
- Quantum optimization is effective.
- Entanglement increases success probability.

J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1

# Results: BiNNs applied to a Gaussian dataset



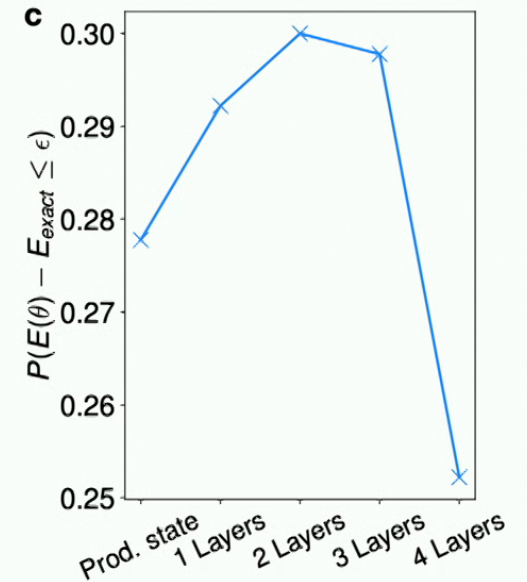
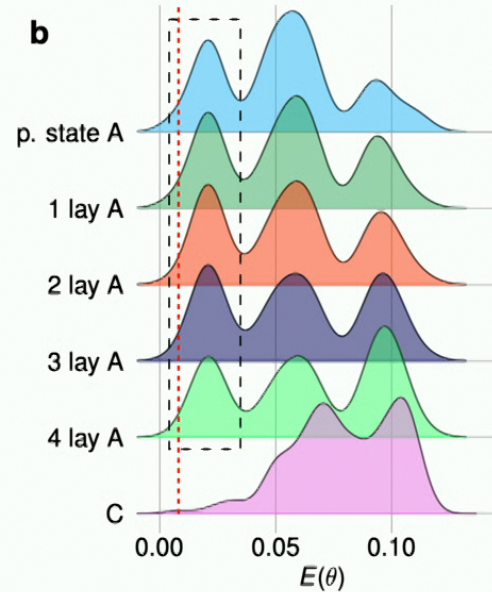
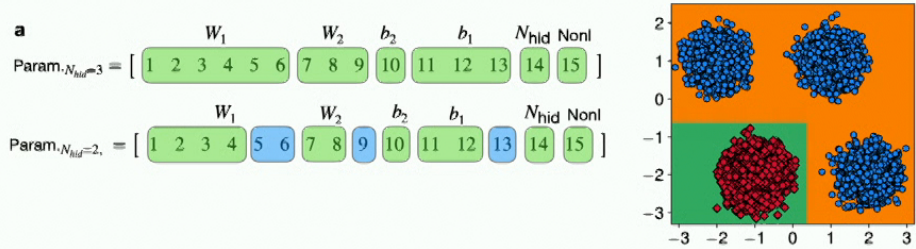
$$\text{Param}_{N_{hid}=3} = \begin{bmatrix} W_1 & W_2 & b_2 & b_1 & N_{hid} & Nonl \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \end{bmatrix}$$

$$\text{Param}_{N_{hid}=2} = \begin{bmatrix} W_1 & W_2 & b_2 & b_1 & N_{hid} & Nonl \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \end{bmatrix}$$

J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1



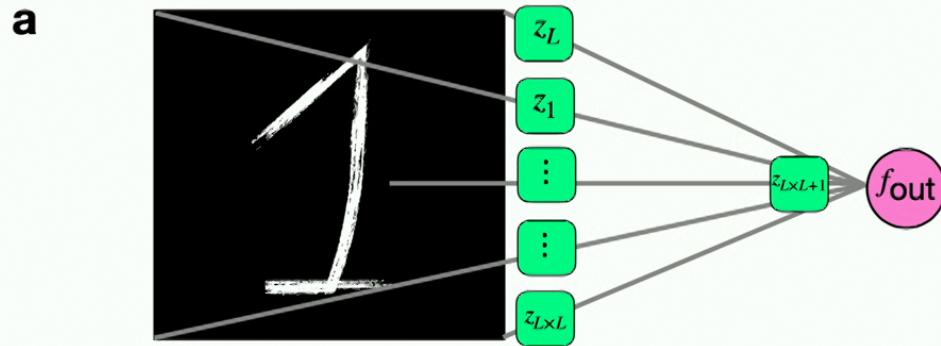
# Results: BiNNs applied to a Gaussian dataset



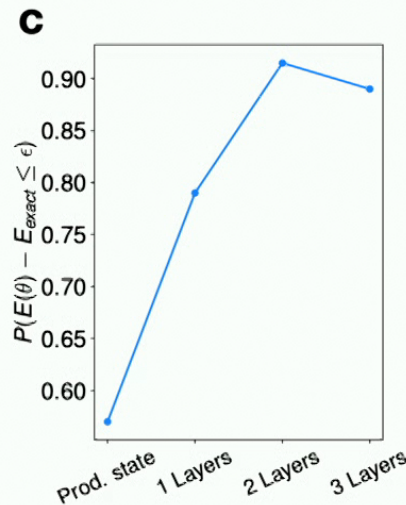
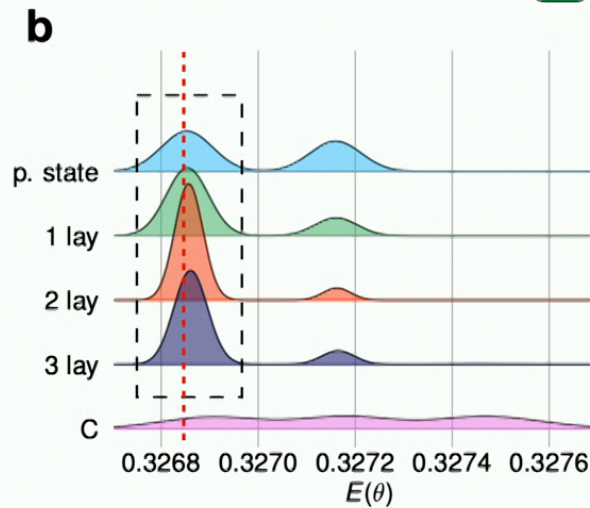
- **Task:** Binary classification of data points
- **Training objective:** Optimize weights, biases, architectural choice of nonlinearity, and hidden layer width (2 or 3).
- Lower success probability but overall successful optimization.
- Optimal circuit depth— optimal use of entanglement

J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1

## Results: BiNNs applied to reduced MNIST (4x4 pixels) dataset



- **Task:** Logistic regression
- **Training objective:** Optimize weights and biases (-3 and -1)



- High success probability.
- Optimal circuit depth— optimal use of entanglement

J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1

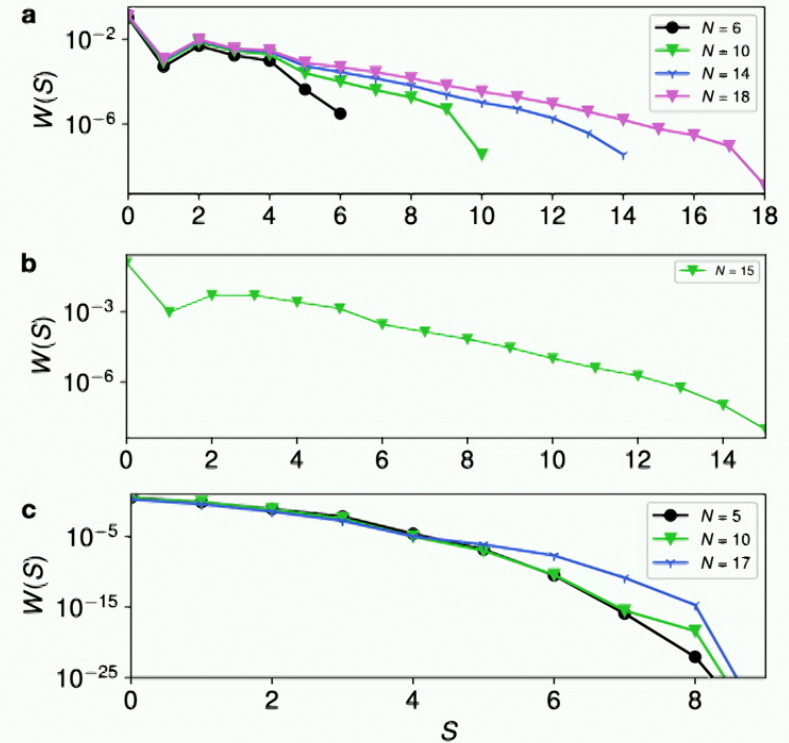


## Results: Global vs Local Objective function

$$\hat{C} = \sum_{\hat{\sigma}_1, \dots, \hat{\sigma}_N} f(\hat{\sigma}_1, \dots, \hat{\sigma}_N) \bigotimes_{i=1}^N \hat{\sigma}_i$$

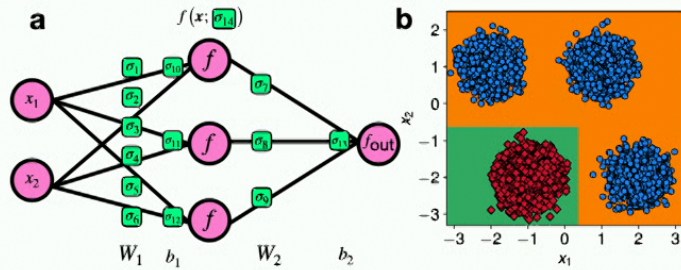
$$W(S) = \sum_{\hat{\sigma}_1, \dots, \hat{\sigma}_N} |f(\hat{\sigma}_1, \dots, \hat{\sigma}_N)|^2 \delta_{S, S(\hat{\sigma}_1, \dots, \hat{\sigma}_N)}$$

- a) and b) are Gaussian datasets with activation and # of layers choice respectively. Dominant contributions are 2- and 3- local.
- c) is the rescale MNIST dataset. Dominant contributions are independent local fields.
- The loss function is predominantly local.

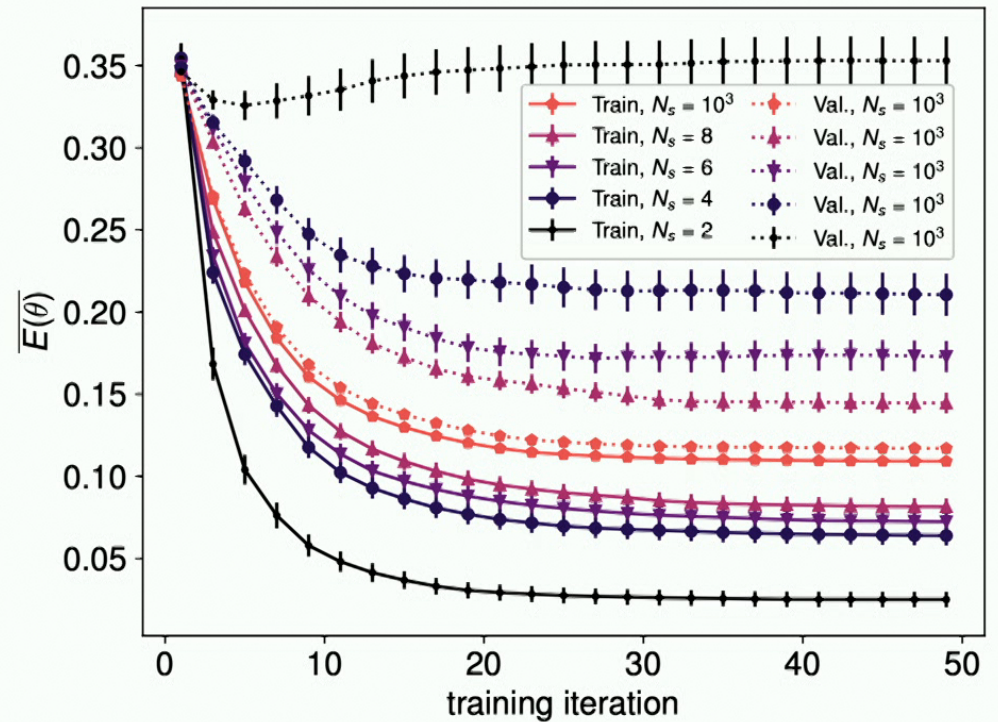


J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1

# Results: Benchmarking single-loop optimization



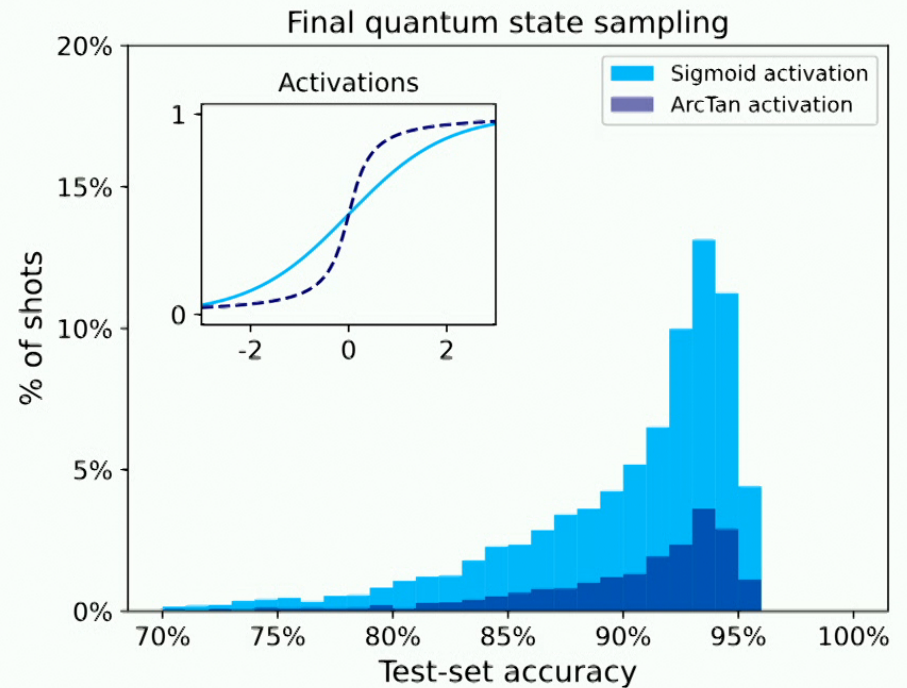
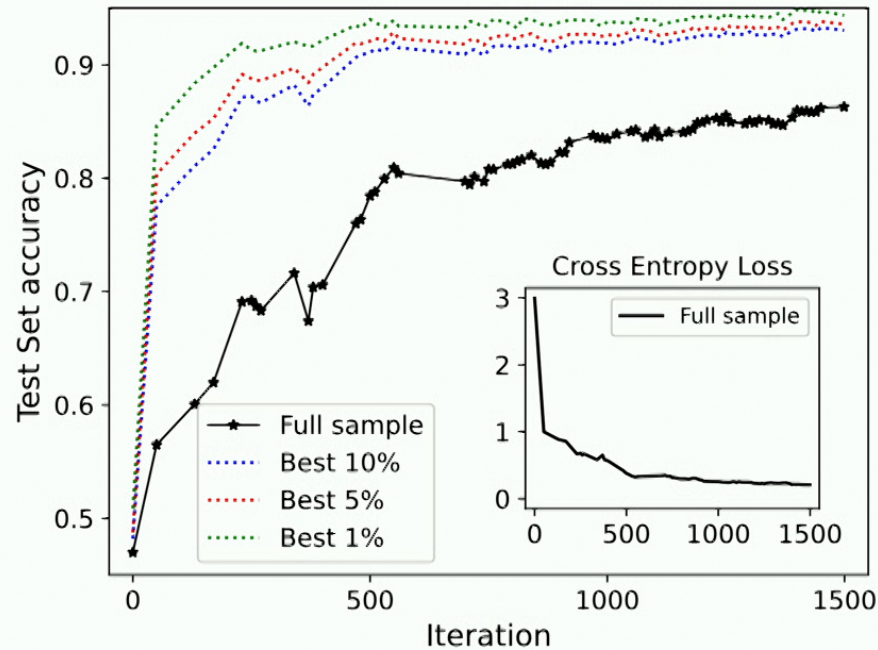
- **Task:** Binary classification of data points
- **Training objective:** Optimize weights, biases and architectural choice of nonlinearity
- $E(\theta)$  is averaged over 100 independent realizations of training datasets of size  $N_s$
- The generalization gap behaves as in standard ML approaches



J. Carrasquilla, M. Hibat-Allah, E.M.I, A. Makhzani, K. Neklyudov, G. W. Taylor, and G. Torlai, arXiv:2301.08292v1



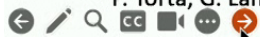
## Preliminary Results: BiNNs applied to reduced MNIST (8x8 pixels) dataset



- **Task:** Logistic regression
- **Training objective:** Optimize weights, biases (-1 and +1) and choice of activation function

- Nqubits=66 – way beyond full enumeration
- Training shots=1500, Validation shots=10,000
- Circuit depth=2, Trained with MPS using QN-SPSA.

P. Torta, G. Lami, J. Carrasquilla, M. Collura, EMI, work in progress



# Conclusions

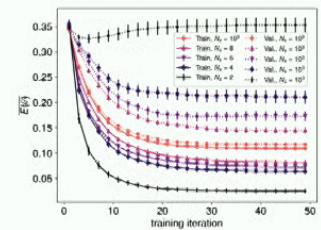
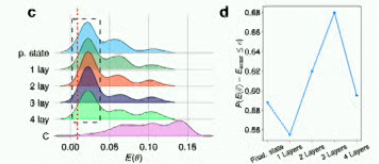
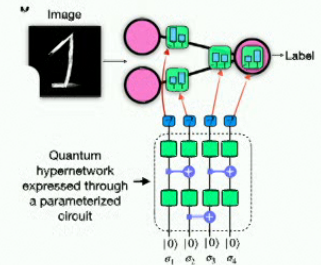
Demonstrated training of BiNNs using variational quantum circuits

Quantum effects such as entanglement were shown to enhance the optimization process

Training, hyperparameters and architectural searches were successfully combined in a single optimization loop

Training BiNNs on quantum computers would help reduced carbon footprint.

Substantial work is still needed to scale our approach to Deep Learning models, e.g multibasis encoding or adaptive layer learning

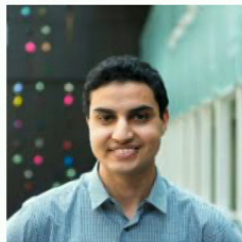




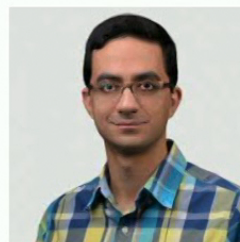
# Thank You!



Juan Carrasquilla  
(Vector institute/UW)



Mohamed Hibat-Allah  
(Vector institute/UW)



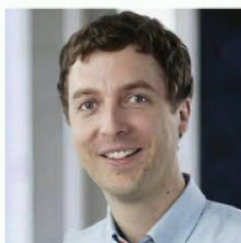
Alireza Makhzani  
(Vector institute/  
UofT)



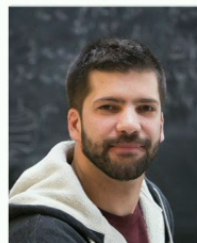
Pietro Torta  
(SISSA)



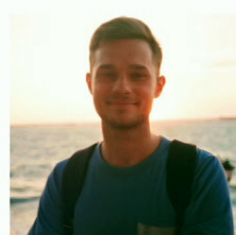
Guglielmo Lami  
(SISSA)



Graham Taylor  
(Vector institute/  
University of Guelph)



Giacomo Torlai  
(AWS)



Kirill Neklyudov  
(Vector institute)



Mario Collura  
(SISSA)