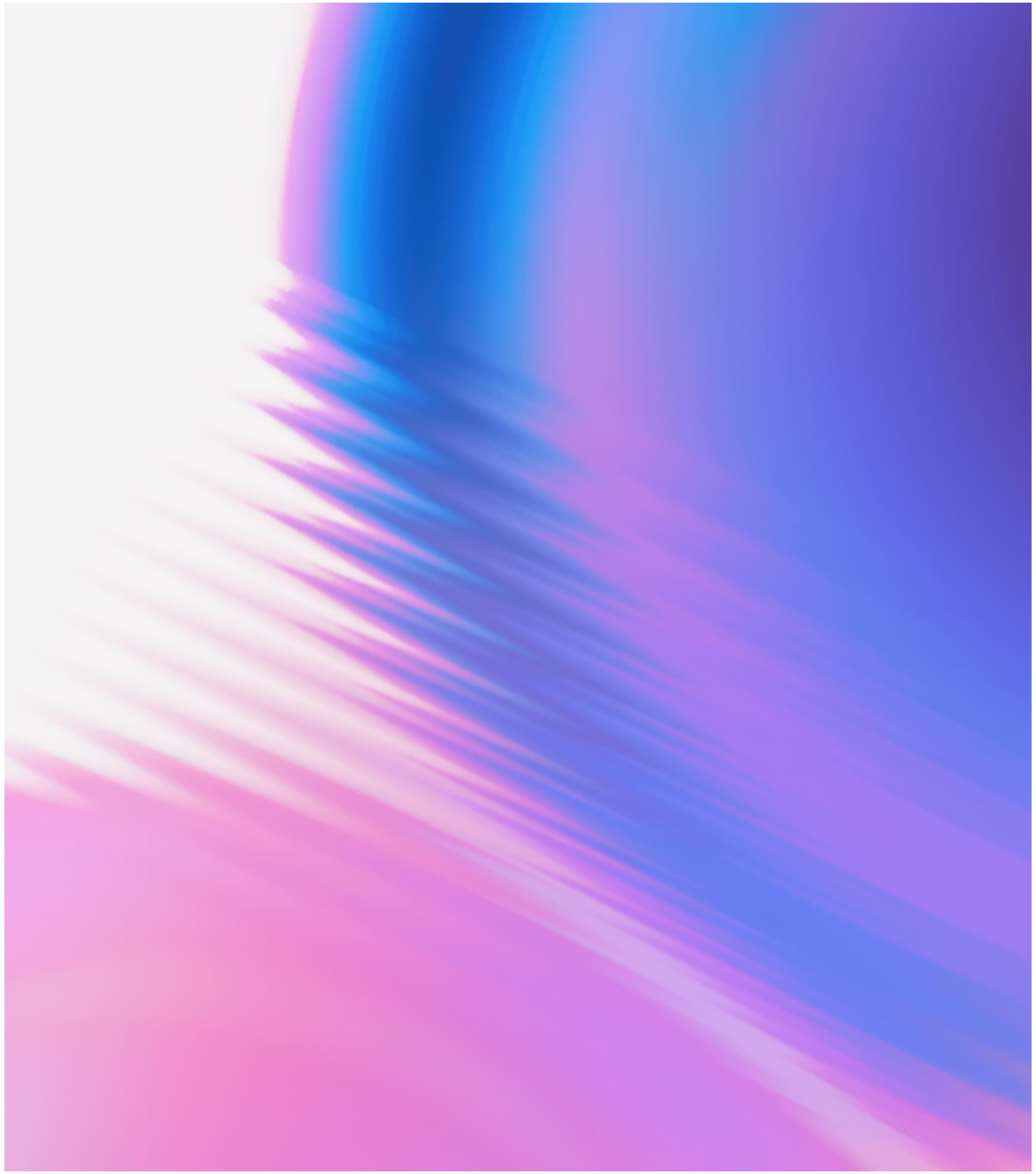
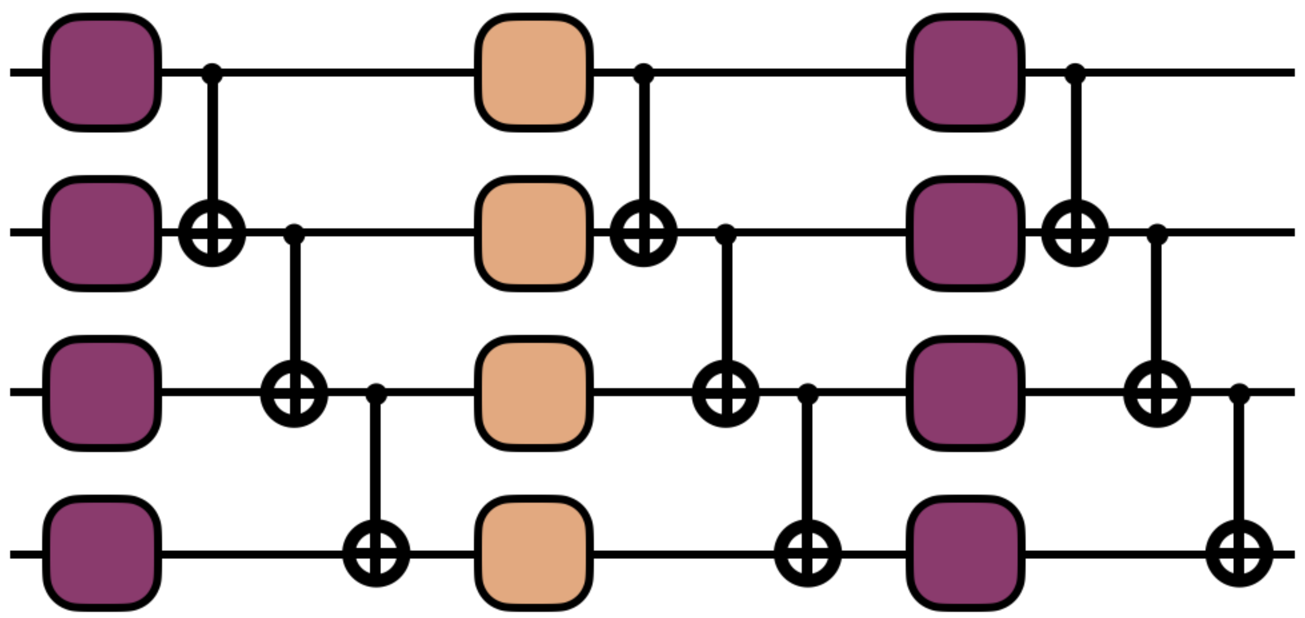
Quantum Neural Networks Diagnostics with Neural Tangent Kernels

Francesco Tacchino
IBM Quantum, IBM Research Europe – Zurich

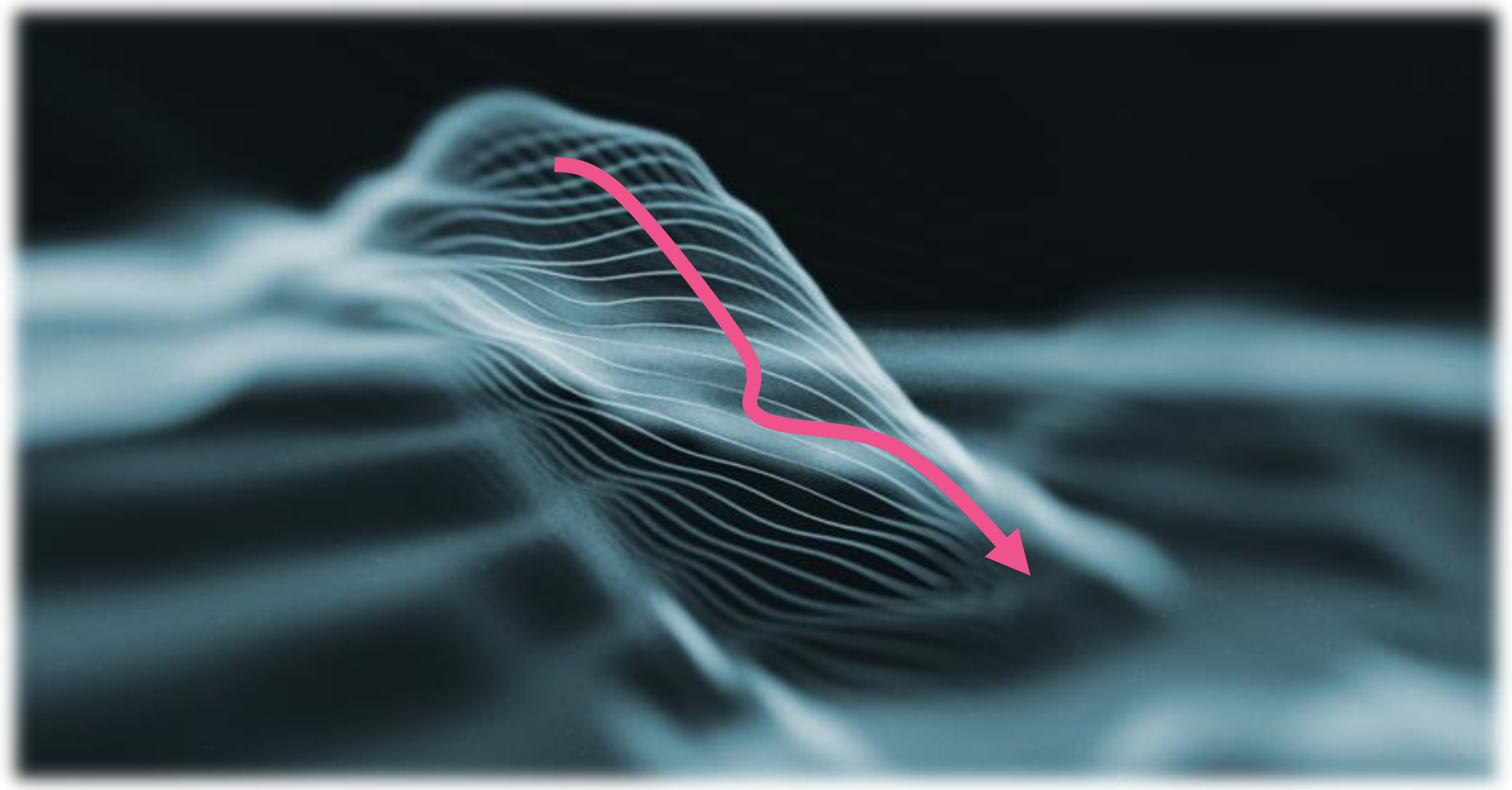


Why do (Q)NNs compute a certain function?

Parametrized quantum circuit

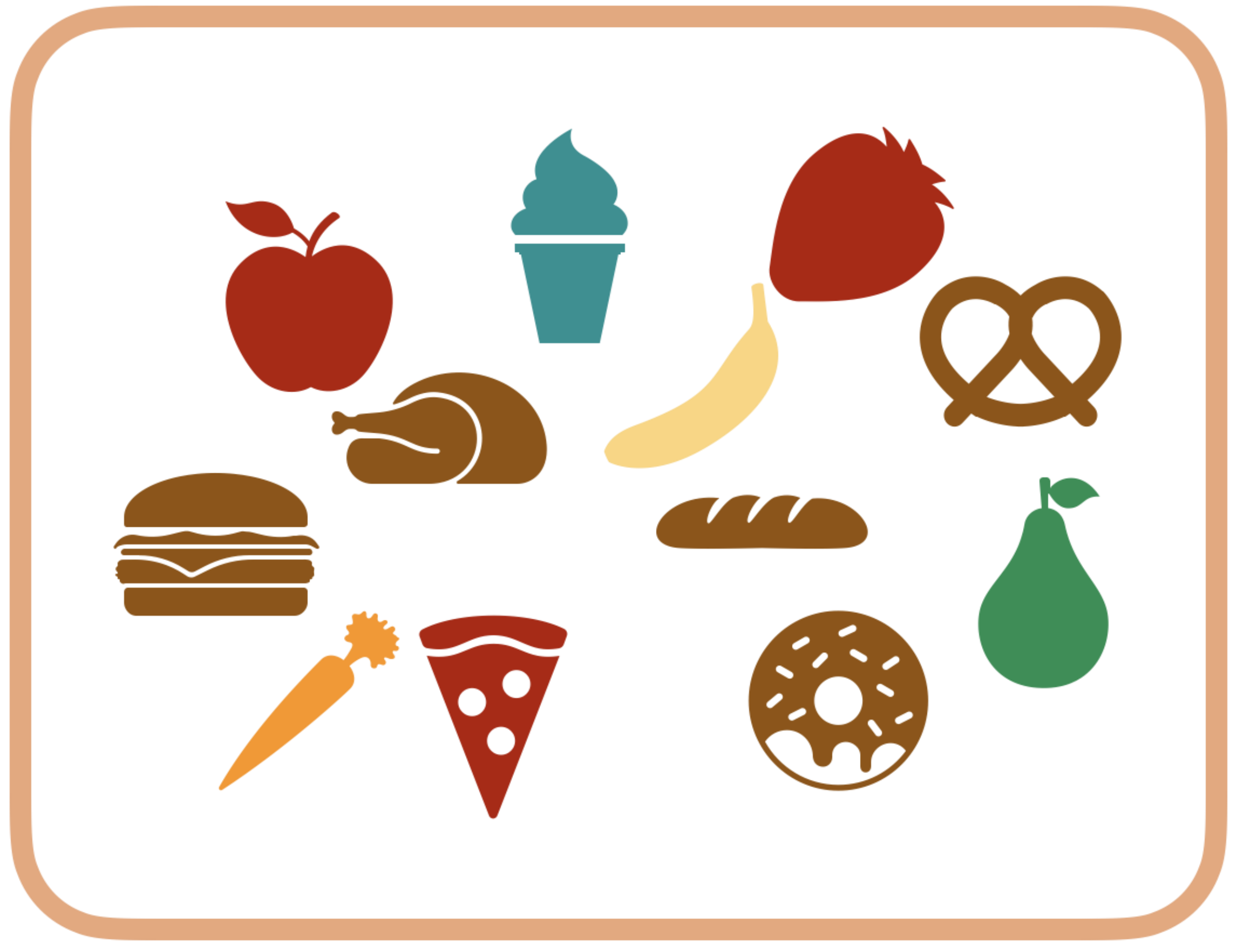


$$f(x; \bar{\theta})$$



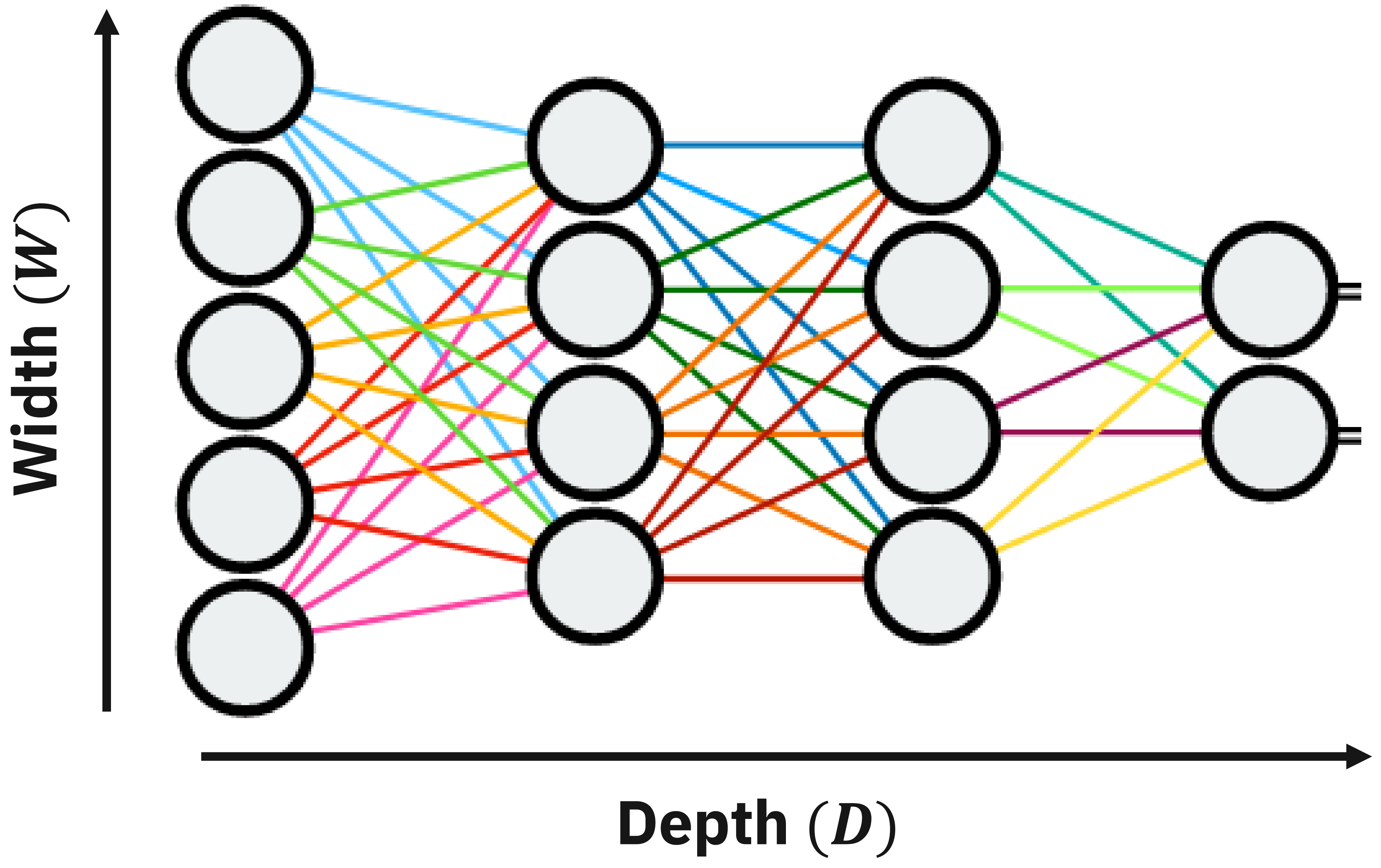
Training algorithm

Dataset



- We know quite well *how* QNNs compute a certain output function (*microscopic model*).
- An effective, *macroscopic* theory of QNNs would be useful to understand the training dynamics, the effect of data and the influence of size and structure of the model on its performances.

The classical case: effective theory of deep learning



Neural Tangent Kernel

$$H_{i_1 i_2;\alpha_1\alpha_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i_1;\alpha_1}}{d\theta_\mu} \frac{dz_{i_2;\alpha_2}}{d\theta_\nu}$$

$$f(x;\bar{\theta}) = f(x;\theta_0) + \frac{df}{d\theta}(\bar{\theta}-\theta_0) + \frac{d^2f}{d\theta^2}(\bar{\theta}-\theta_0)^2 + \ldots$$

- **Infinite width limit $(W \to \infty)$**

$$\lim_{W\to\infty} \bar{\theta} = [\bar{\theta}]\left(\theta, f, \frac{df}{d\theta}; \mathscr{D}\right)$$

Training data

Concentrated ("frozen") Neural Tangent Kernel

- **Perturbation theory around $1/W \to 0$** – at first order

$$\lim_{W\to\infty} \bar{\theta} = [\bar{\theta}]\left(\theta_0, f, \frac{df}{d\theta}, \frac{d^2f}{d\theta^2}, \frac{d^3f}{d\theta^3}; \mathscr{L}; \mathscr{D}\right)$$

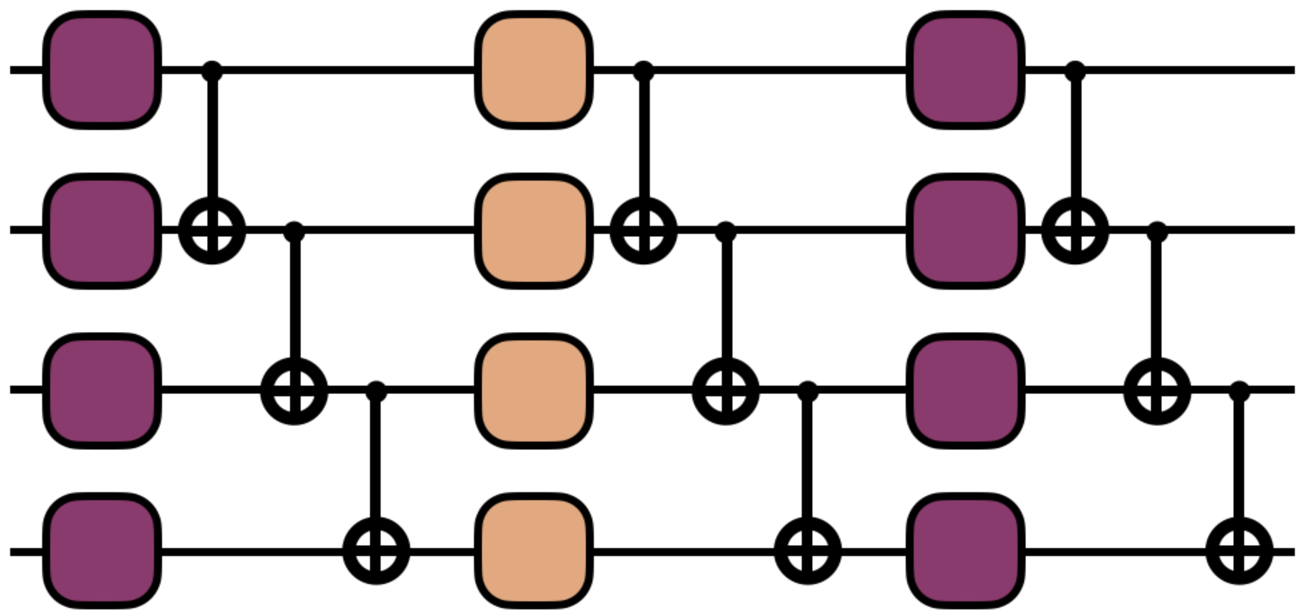
Learning algorithm

Closed formula for finite width-NN prediction (fits on a T-shirt [1])

[1] Roberts, Yaida, Hanin, *The Principles of Deep Learning Theory*

The Quantum Neural Tangent Kernel

IBM **Quantum**



- Consider a parametrized quantum circuit model and a quadratic loss function:

$$U(\vec{\theta}) = \prod_{\ell=1}^{L} W_\ell \exp\left(i\theta_\ell X_\ell\right) \equiv \prod_{\ell=1}^{L} W_\ell U_\ell \qquad \mathscr{L} = \frac{1}{2}\sum_{i=1}^{N}\left(f(x_i,\theta) - y_i\right)^2 = \frac{1}{2}\sum_{i=1} \varepsilon_i^2$$

- Apply gradient descent with learning rate η:

$$\delta\theta_l = \theta_l(t+1) - \theta_l(t) = -\eta\frac{\partial\mathscr{L}}{\partial\theta_l} = -\eta\sum_i \epsilon_i \frac{\partial\epsilon_i}{\partial\theta_l}$$

$$\delta\epsilon_i \simeq \sum_l \frac{\partial\epsilon_i}{\partial\theta_l}\delta\theta_l = -\eta\sum_l\sum_{i'}\frac{\partial\epsilon_i}{\partial\theta_l}\frac{\partial\epsilon_{i'}}{\partial\theta_l}\epsilon_{i'} = -\eta\sum_{i'} K_{ii'}\epsilon_{i'}$$

QNTK

$$K_{ii'} = \sum_l \frac{\partial\varepsilon_i}{\partial\theta_l}\frac{\partial\varepsilon_{i'}}{\partial\theta_l}$$

The lazy training, frozen QNTK regime

If, for any reason, the variational parameters do not change much during training (*lazy training*), the QNTK remains essentially constant and the model behaves as its linearization.

$$\left.\begin{array}{l} \dfrac{\|\theta^{(0)} - \theta^{(t)}\|}{\|\theta^{(0)}\|} \ll 1 \\[2ex] \forall k > 1 \quad \left.\dfrac{\partial^k f}{\partial\theta^k}\right|_{\theta=\theta_0} (\theta-\theta_0)^k \to 0 \end{array}\right\} \quad K \simeq \mathbb{E}_\theta[K]$$

QNTK – analytical results

Under the frozen QNTK assumption, we can prove that the **residual training error decays exponentially**:

$$\epsilon_i(t) = \sum_{i'} [e^{-\eta K t}]_{ii'} \epsilon_{i'}(0)$$

To get to this result, we essentially need:

- small learning rate η

More on this later

- concentrated QNTK

$$K \simeq \mathbb{E}_\theta[K] = \bar{K}$$

$$\frac{\Delta K}{\bar{K}} \ll 1$$

Fluctuations

- **Formal analysis: overparametrization (large D limit)**
  Under a 2-design assumption for the ansatz, we have

$$\bar{K} \approx \frac{L\mathrm{Tr}\left(O^2\right)}{D^2}$$

Hilbert space dimension!

For the decay rate to be $O(1)$ we need $L \approx D$ or $L \approx D^2$.

- **Formal analysis: concentration conditions (large D limit)**
  Under a 4/6-design assumption for the ansatz, we have

$$\frac{\Delta K}{\bar{K}} \approx \frac{1}{\sqrt{L}} \ll 1 \qquad \frac{\Delta \mu}{\bar{K}} \approx \frac{\left(\eta\sqrt{\mathrm{Tr}(O^2)}\right)}{D} \ll 1$$

$$\mu = \sum_{\ell_1,\ell_2} \frac{\partial^2 \varepsilon}{\partial\theta_{\ell_1}\partial\theta_{\ell_2}} \frac{\partial \varepsilon}{\partial\theta_{\ell_1}} \frac{\partial \varepsilon}{\partial\theta_{\ell_2}}$$

The quantum meta-kernel (dQNTK) controls higher order contributions in η

[J. Liu et al., PRX Quantum 3, 030323 (2022); J. Liu et al., Phys. Rev. Lett. 130, 150601 (2023)]

Main formula for QNTK diagnostics

How can the QNTK be used *in practice* to gain insights on actual models, gauge their performances and guide their design?

Under the frozen QNTK assumption, we can write an analytic formula for the output of a QNN model at any training step t, i.e. we can compute the prediction that the model would make on an input x_j taken from the data disribution, **using only properties of the model at initialization**

$$\mathbb{E}_\theta[f(x_j, \theta_t)] = \sum_{i',\alpha,\gamma} \mathbb{E}_\theta\left[K_{j,i'}[K^{-1}]_{i',\alpha}\left[\mathbb{I} - e^{-\eta K t}\right]_{\alpha,\gamma}\right] y_\gamma$$

These indices run only over the elements of the training set

Training diagnostic

$x_j \in training\ set$, predictions as function of t

- Training speed
- Critical learning rate

Inference diagnostic

$x_j \in test\ set,\ t \to \infty$

- True risk estimation & overfitting

Training dynamics

We can get useful insights on the training properties of the model by analyzing the **QNTK eigenspectrum at initialization**

- **Vector notation.** To understand how, let us first adopt a vector notation over the elements of the data set:

$$\mathbb{E}_\theta[f(\mathcal{X},\boldsymbol{\theta}_t)] = \begin{bmatrix} \mathbb{E}_\theta[f(\mathbf{x}_0,\boldsymbol{\theta}_t)] \\ \vdots \\ \mathbb{E}_\theta[f(\mathbf{x}_M,\boldsymbol{\theta}_t)] \end{bmatrix} \qquad \mathcal{Y} = \begin{bmatrix} y_0 \\ \vdots \\ y_M \end{bmatrix}$$

- **QNTK diagonalization.** Let us also assume that the QNTK matrix can be diagonalized as follows:

$$K = A^T D A \qquad D = \mathrm{diag}\left(\lambda_0, \ldots \lambda_{M-1}\right)$$

$$A e^{-\eta K t} A^T = e^{-\eta D t}$$

$$\widetilde{\mathbb{E}}_\theta[f(\mathcal{X},\boldsymbol{\theta}_t)] = A\mathbb{E}_\theta[f(\mathcal{X},\boldsymbol{\theta}_t)]$$

$$\tilde{\mathcal{Y}} = A\mathcal{Y}$$

- **Training dynamics.** We can now rewrite the evolution of the mean residual training errors as

$$\widetilde{\mathcal{E}}_i(t) = \mathbb{E}_\theta\left[\mathbb{I} - e^{-\eta\lambda_i t}\right]\tilde{\mathcal{Y}}_i - \tilde{\mathcal{Y}}_i$$

- **Result 1: critical learning rate.** Since $e^{-\eta\lambda_i t} \approx (1-\eta\lambda_i)^t$, we can ensure that the training does not diverge by setting

$$|1-\eta\lambda_i| < 1 \implies \eta_{crit} \approx 2/\lambda_{max}$$

- **Result 2: training decay time.** Using the critical learning rate, the slowest component converges as

$$\widetilde{\mathcal{E}}_{min}(t) = \mathbb{E}_\theta\left[\left(\mathbb{I} - e^{-2\frac{1}{\kappa}t}\right)\tilde{\mathcal{Y}}_{min}\right] - \tilde{\mathcal{Y}}_{min}$$

QNTK condition number $\kappa = \lambda_{\max}/\lambda_{\min}$

$$\eta_{crit}\lambda_{min} = 2\kappa^{-1} = \tau^{-1}$$

Inference

The long time limit of the frozen QNTK prediction formula gives us insights about the performances that a fully trained QNN would have.

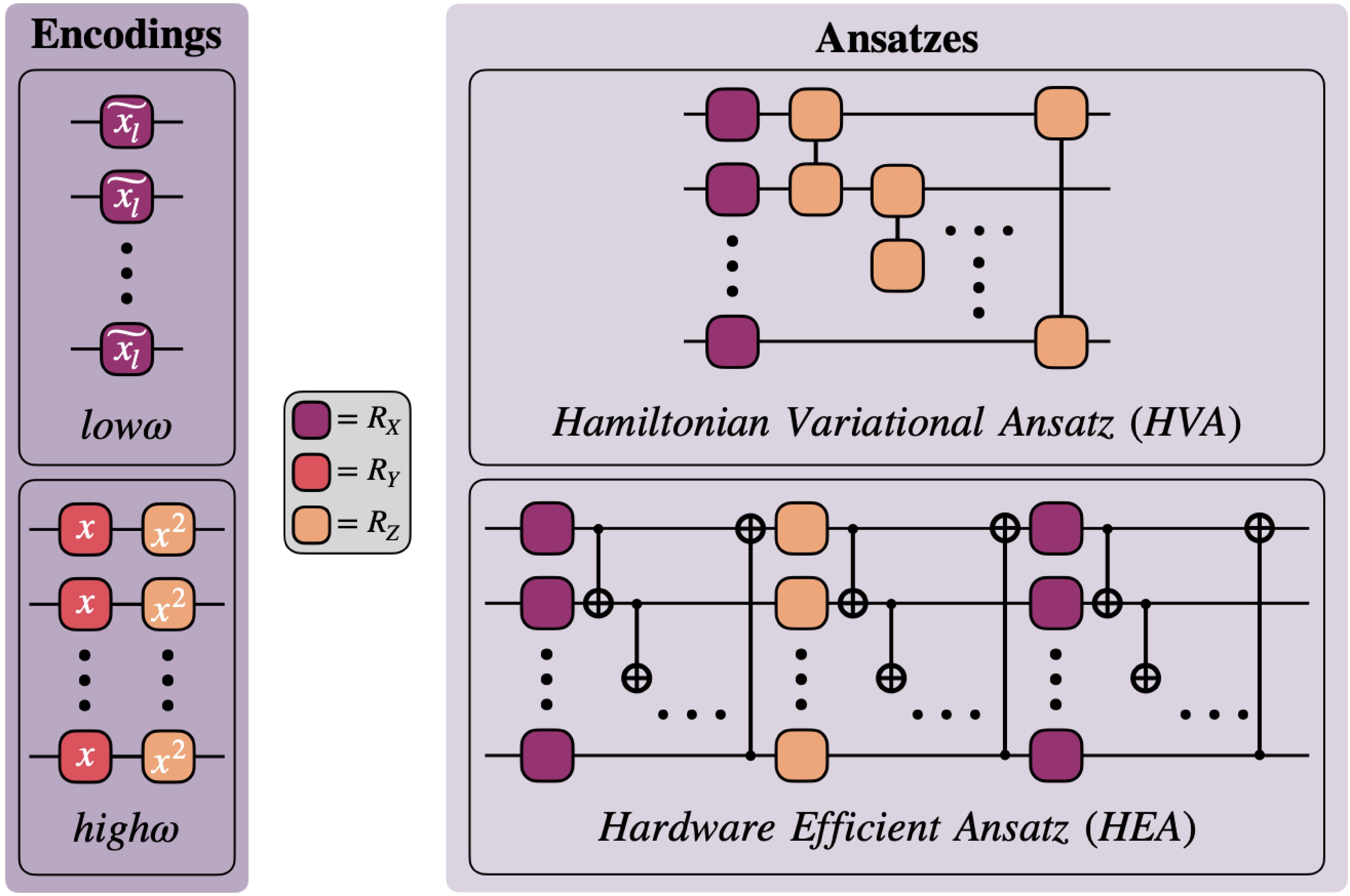
$$\mathbb{E}_{\theta}[f(x_j, \theta_t)] = \sum_{i',i} K_{j,i'}[K^{-1}]_{i',i} y_i$$
$$t \to \infty$$

Notice that:

- This is a linear model prediction, where the QNTK plays the role of the embedding kernel
- By definition, exact labels are predicted for all training points

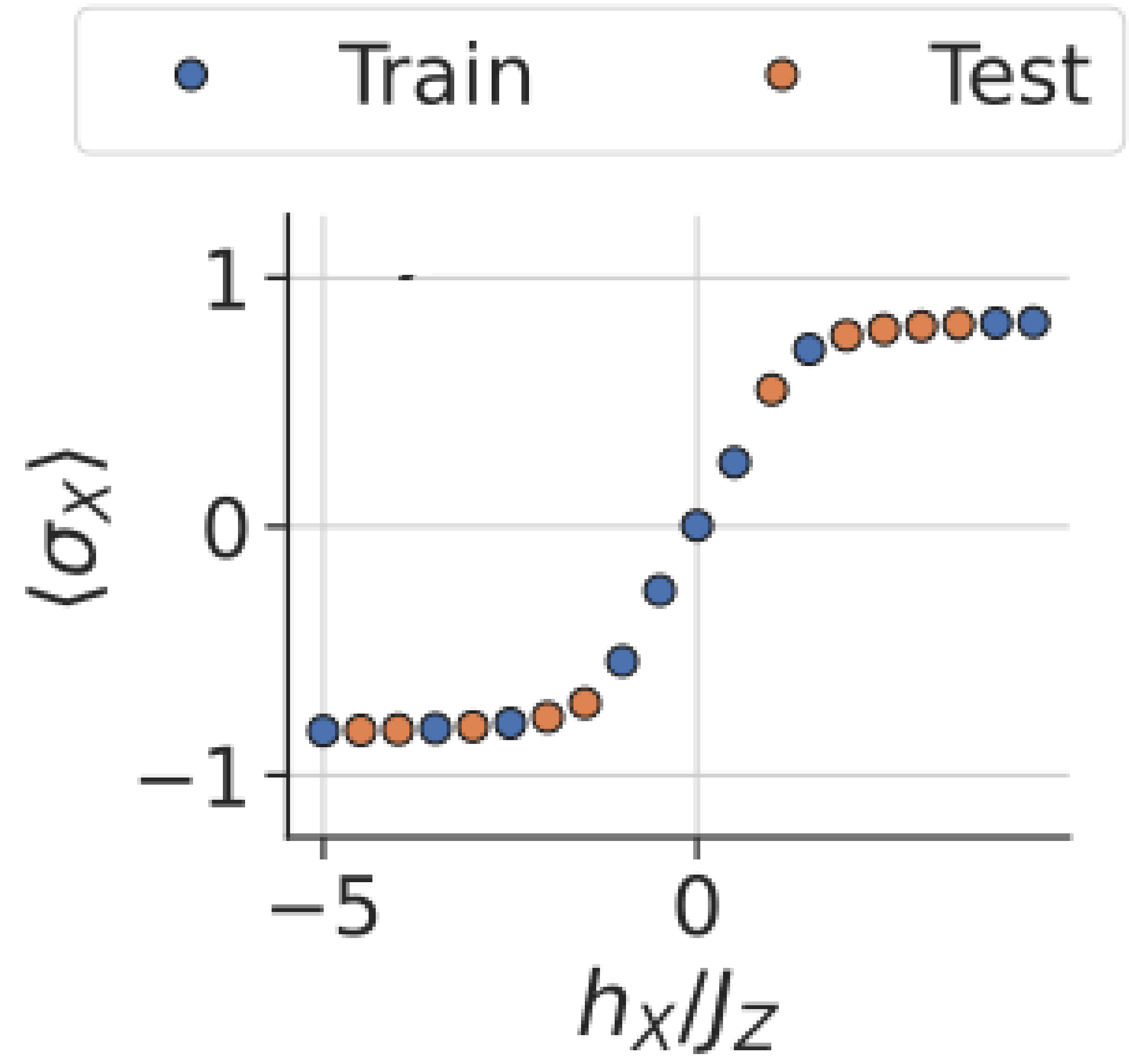
IBM **Quantum**

Numerical experiments

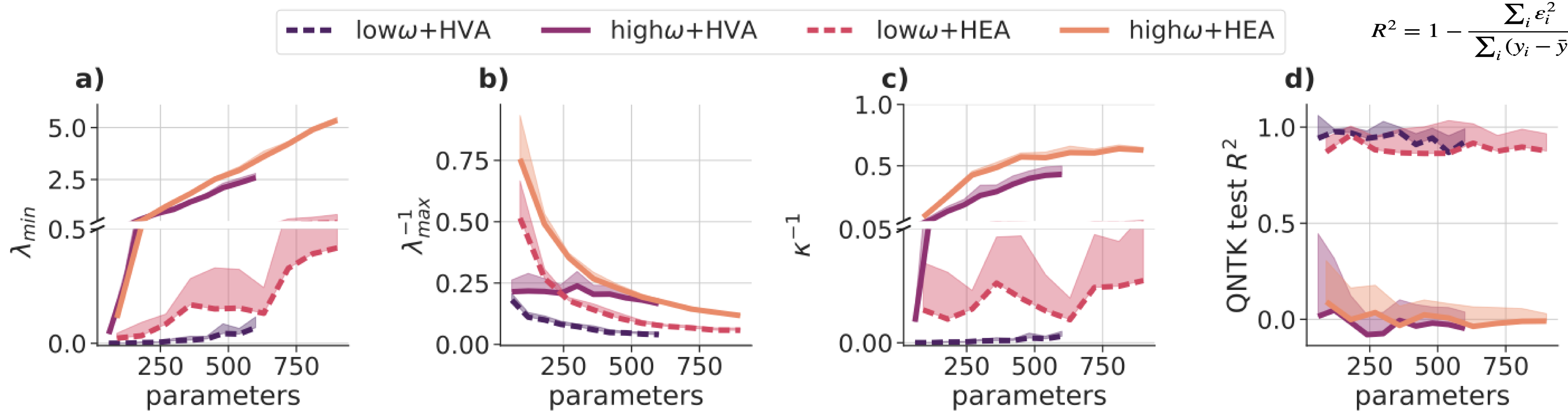


Dataset: magnetization of the 1D transverse field Ising model as a function of the field strength ($N = 6$)

$$H = -J_Z \sum_{i=1}^{N} \sigma_Z^i \sigma_Z^{i+1} - h_X \sum_{i=1}^{N} \sigma_X^i$$



Results (I) – QNTK features

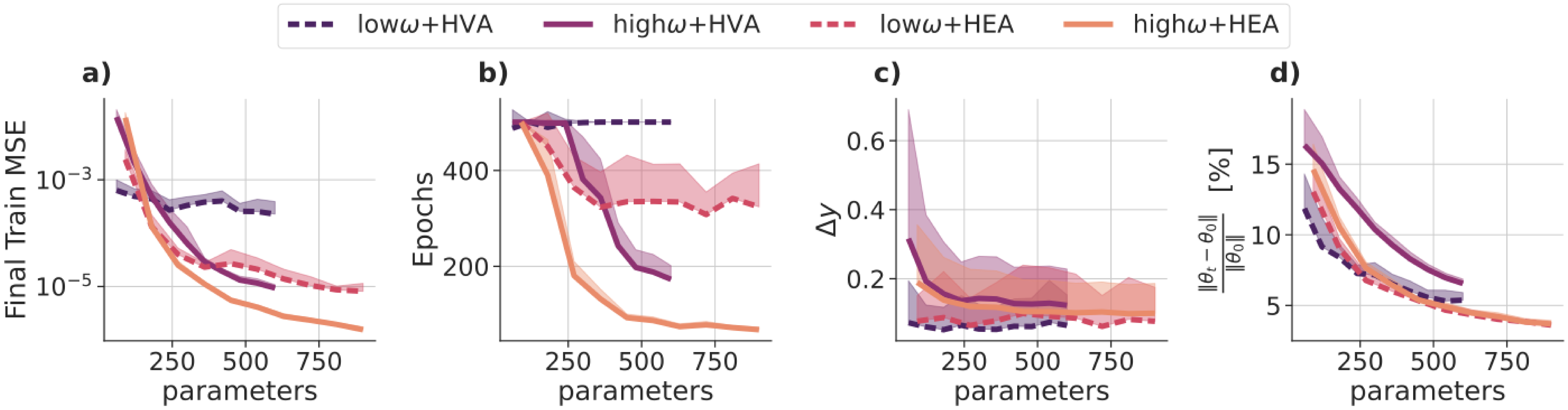


$$R^2 = 1 - \frac{\sum_i \varepsilon_i^2}{\sum_i (y_i - \bar{y})^2}$$

- $\lambda_{\min}$ grows with number of parameters ⇒ faster training
- high-ω and HEA are expected to train faster than low-ω and HVA
- Similar hierarchy for the maximum allowed learning rate, which is proportional to $\lambda_{\max}^{-1}$
- Overall, the characteristic training time $\tau \propto \kappa$ decreases with more parameters
- Low condition numbers seen for high-ω models may indicate overfitting. This is confirmed by the R^2 scores on the training set for QNTK predictions.

Results (II) – Training at $\eta = \eta_{\text{crit}}$

IBM **Quantum**

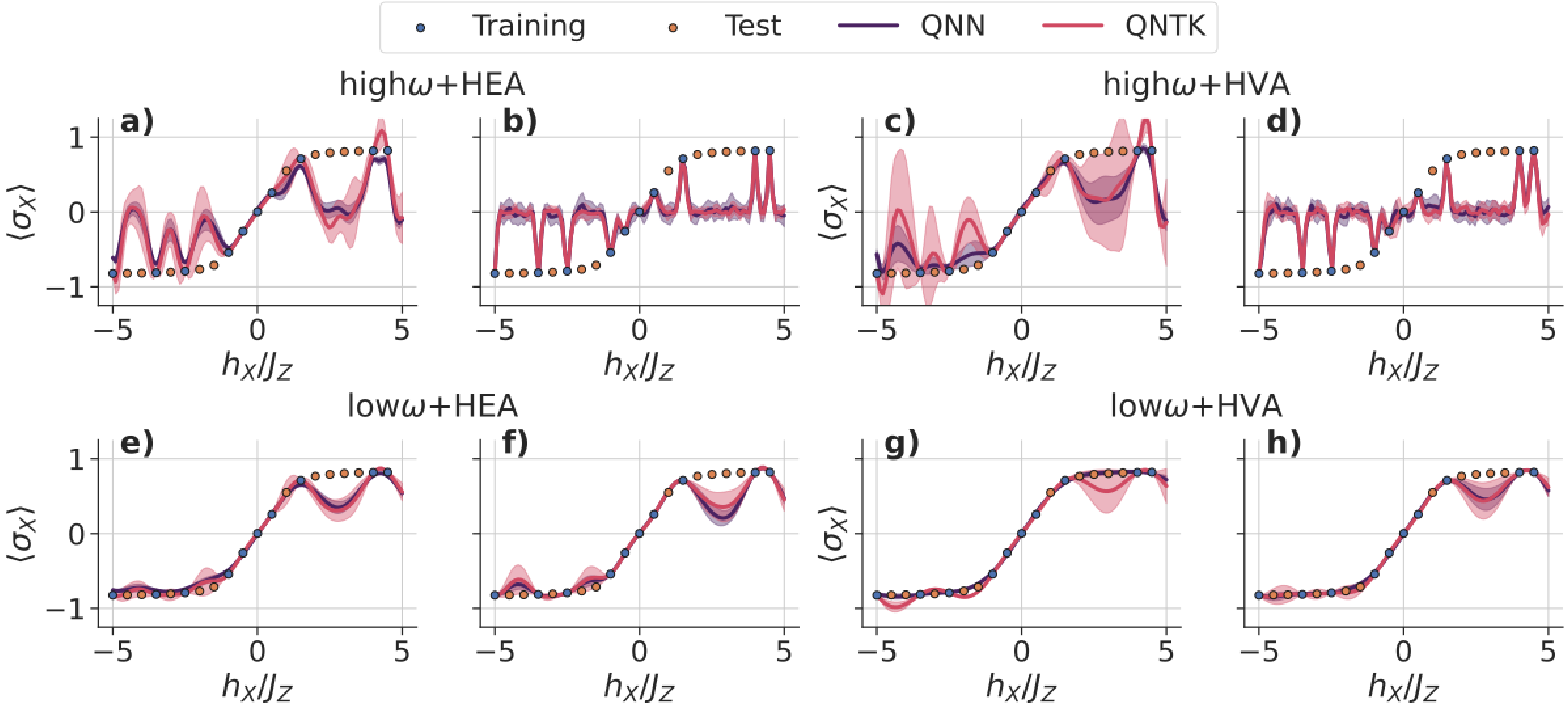


- Confirming our predictions, high-ω and HEA train faster and to lower MSE than than low-ω and HVA.
- The QNTK predictions are close to the actual results from the training model, as quantified by the average absolute difference. The distance decreases consistently as we approach a true *lazy training* regime.

$$\Delta y = \frac{1}{D} \sum_{i}^{D} \left| y_i^{QNN} - y_i^{QNTK} \right|$$

Results (III) – Inference at $\eta = \eta_{\text{crit}}$

IBM **Quantum**



Summary

IBM **Quantum**

- **Effective theory for QNNs**
  - Definition of QNTK
  - Provable training convergence for overparametrized, wide models
  - Formal analysis of concentration conditions for the frozen QNTK limit
  - [not in this talk] Differential QNTK limit

J. Liu, F. Tacchino, J.R. Glick, L. Jiang, A. Mezzacapo, *Representation learning via quantum neural tangent kernels*, PRX Quantum 3, 030323 (2022)

J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, A. Mezzacapo, *Analytic theory for the dynamics of wide quantum neural networks,* Phys. Rev. Lett. 130, 150601 (2023)

- **QNTK diagnostics**
  - Average model prediction formula
  - Critical learning rate & characteristic training time
  - Inference predictions
  - Numerical study on realistic datasets

With F. Scala, C. Zoufal and D. Gerace
(manuscript in preparation)

IBM **Quantum**