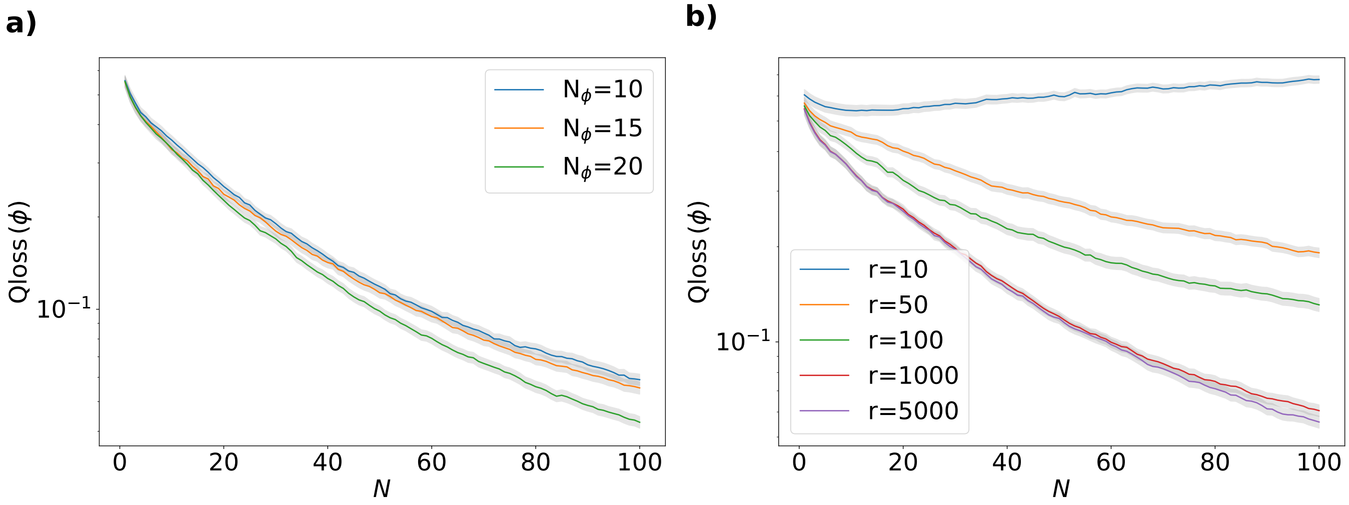
Supplementary Materials for Deep reinforcement learning for quantum multiparameter estimation

Valeria Cimini,[1] Mauro Valeri,[1] Emanuele Polino,[1] Simone Piacentini,[2,3] Francesco Ceccarelli,[3] Giacomo Corrielli,[3] Nicolò Spagnolo,[1] Roberto Osellame,[3] and Fabio Sciarrino[1,*]

[1] *Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 5, I-00185 Roma, Italy*
[2] *Dipartimento di Fisica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, I-20133 Milano, Italy*
[3] *Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche (IFN-CNR),*
Piazza Leonardo da Vinci, 32, I-20133 Milano, Italy

Supplementary Note 1. Effects related to the finite-size of the Bayesian neural network training set.

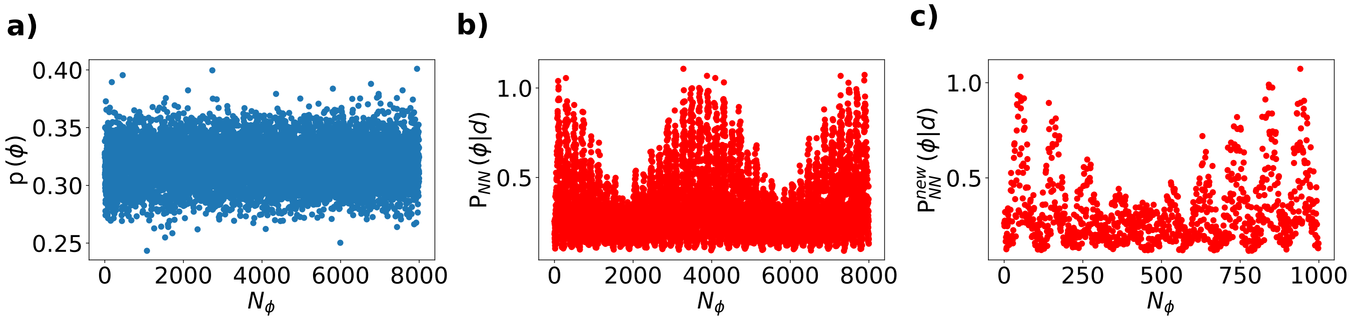
The reconstruction of the single-measurement posterior distribution done by the neural network (NN) for Bayesian update will depend on the number of points in the training grid. Given the complexity of our multiparameter problem we inspect the results achieved with $N_\phi = 10, 15, 20$, corresponding to a total of $1000, 3375, 8000$ grid points respectively in the interval $[0, \pi]$. The overall performances will depend also on the number of repetitions r of the single-measurement results per grid points present in the training set. The dependence on both N_ϕ and r of the performances achieved in terms of Qloss on simulated data are reported in Supplementary Fig.1. The total number of grid points will also determine, as discussed in the main text, the number of particles used in the Bayesian protocol. For what concerns the experiment, in order to implement the adaptive protocol we choose to collect a grid of 8000 points in the extended region of $[-\pi, \pi]$, to have the possibility of applying feedbacks in the interval $[-\pi, 0]$. Such a choice is at the origin of the difference among the results achieved with the Bayesian update done through the explicit system's model (red line) and the one computed via the NN probabilities (magenta line) reported in panel b) of Fig.6 of the main text.



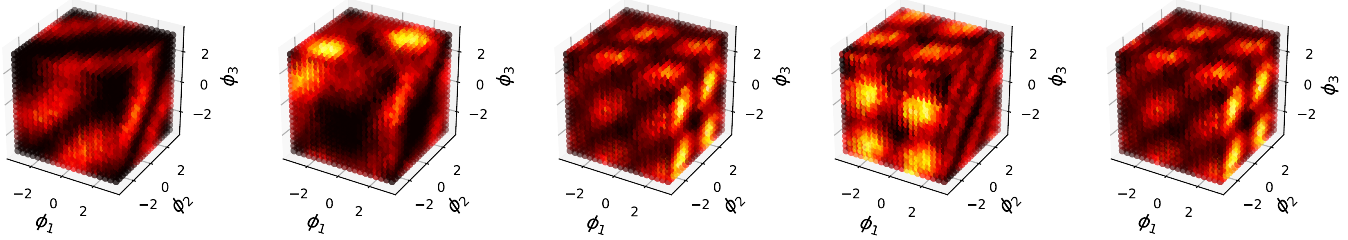
Supplementary Figure 1. Qloss as a function of the number of probes N interacting with the investigated system. **a)** Results achieved with training set of different size. **b)** Results obtained with different measurements statistic.

In this scenario the prior and consequently the single-measurement posterior distribution learned by the NN are defined in the 2π interval and are reported in Supplementary Fig.2 and Fig3. However, for the estimation procedure we have to consider only the $[0, \pi]$ interval. Therefore, at each step of the protocol we first shift the reconstructed *a-priori* and posterior distributions depending on the set values of the control phases and then we cut the shifted distributions in the interval of interest. The Bayesian update will be done with the reduced distribution using therefore $8000/2^3$ particles.

[*] fabio.sciarrino@uniroma1.it



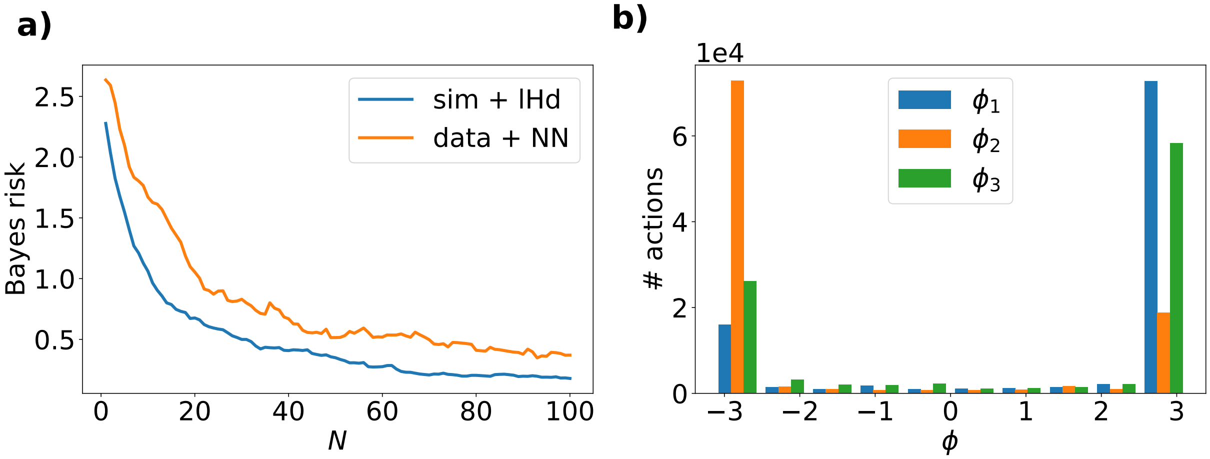
Supplementary Figure 2. **a)** Prior distribution reconstructed for each data point of the training set.**b)** Bayesian posterior NN distribution for the outcome $d = 6$. **c)** Posterior probability used for the Bayesian update, relative to the outcome $d = 6$, in the reduced interval $\phi \in [0, \pi]$.



Supplementary Figure 3. Experimental posterior probability distributions reconstructed by the NN for the outcomes $d = 2, 4, 6, 8, 9$.

Supplementary Note 2. Training of the reinforcement learning algorithm

The training of the RL agent has been performed through the cross-entropy method analyzing the results of 10^4 different episodes for each of the 100 available probes. The initial weights of the 100 NN selecting the action performed by the agent are sampled from a gaussian distribution initially centered in zero, with standard deviation $\sigma = 0.5$ and fixing the threshold selecting the *elite* weights to the 10% of the overall sample. Once trained we use the NN with the selected *elite* weights to compute the Bayes risk on 100 independent experiments all consisting of 100 different measurements. The results obtained when the RL algorithm is run using the likelihood function for both the Bayesian update and for simulating the measurement outcomes are reported in blue in panel a) of Supplementary Fig.4. Here, the Bayes risk is compared with the one achieved when substituting the explicit posterior with the NN probability learned on the grid of measured data and the experimental results are selected with the respective probability from the same grid (orange line). Due to the presence of experimental noise and to the difference among the transformation implemented by the ideal device from the real one, the discrepancy in the obtained Bayes risk is reflected in the Qloss difference from the magenta experimental points and the continuous curve reported in panel c) of Fig.6. The control feedbacks chosen by the agent on all the episodes of the training are reported in panel b) of Supplementary Fig.4.



Supplementary Figure 4. **a)** Bayes risk as a function of the number of sent probes N. **b)** Histogram of different actions imparted by the RL agent for each experiment consisting of 100 probes. The experiments are performed for 100 different phases values repated 10 times each. The three colors refer to feedback imparted on the three different investigated phases.