Machine-learning-based device-independent certification of quantum networks

Nicola D'Alessandro,[1] Beatrice Polacchi ●,[1] George Moreno,[2,3] Emanuele Polino ●,[1]
Rafael Chaves,[2,4] Iris Agresti ●,[1,5,*] and Fabio Sciarrino[1]

[1]*Dipartimento di Fisica - Sapienza Università di Roma, P.le Aldo Moro 5, I-00185 Roma, Italy*
[2]*International Institute of Physics, Federal University of Rio Grande do Norte, 59078-970, P. O. Box 1613, Natal, Brazil*
[3]*Departamento de Computação, Universidade Federal Rural de Pernambuco, 52171-900 Recife, Pernambuco, Brazil*
[4]*School of Science and Technology, Federal University of Rio Grande do Norte, 59078-970 Natal, Brazil*
[5]*University of Vienna, Faculty of Physics & Research Network Quantum Aspects of Space Time (TURIS),*
Boltzmanngasse 5, 1090, Vienna, Austria

Witnessing nonclassical behavior is a crucial ingredient in quantum information processing. For that, one has to optimize the quantum features a given physical setup can give rise to, which is a hard computational task currently tackled with semidefinite programming, a method limited to linear objective functions and that becomes prohibitive as the complexity of the system grows. Here, we propose an alternative strategy, which exploits a feedforward artificial neural network to optimize the correlations compatible with arbitrary quantum networks. A remarkable step forward with respect to existing methods is that it deals with nonlinear optimization constraints and objective functions, being applicable to scenarios featuring independent sources and nonlinear entanglement witnesses. Furthermore, it offers a significant speedup in comparison with other approaches, thus allowing to explore previously inaccessible regimes. We also extend the use of the neural network to the experimental realm, a situation in which the statistics are unavoidably affected by imperfections, retrieving device-independent uncertainty estimates on Bell-like violations obtained with independent sources of entangled photon states. In this way, this work paves the way for the certification of quantum resources in networks of growing size and complexity.

I. INTRODUCTION

The last decades have witnessed an outstanding development of quantum technologies. The discovery of correlations that are inexplicable by the classical laws of physics has boosted remarkable advances in communication security [1] and computing capabilities [2]. In this context, it is essential to be equipped with techniques to certify the presence of nonclassical features within a given physical process. In particular, this is relevant in quantum communication and cryptography protocols [3–10], self-testing [11], and randomness generation [12]. For this reason, a class of certification protocols, known as device-independent (DI), has been introduced and experimentally tested in different photonic scenarios [13,13–30]. There, conclusions about the operation of a given apparatus are drawn without making assumptions about its inner functioning details [11,31–33]. Indeed, the working principle of DI protocols is only based on the discrepancy between quantum and classical causal predictions. For example, let us consider the Bell causal

structure shown in Fig. 1(a), involving two parties. Here, the cause-effect relationships among the variables (and the laws of classical physics) imply some mathematical constraints on the input-output correlations $p(a, b|x, y)$. In this notation, $p(a, b|x, y)$ is the probability that the two parties respectively get the outcomes labeled as a and b, when performing the measurements operators indicated by x and y. One of such constraints is the Clauser-Horne-Shimony-Holt (CHSH) inequality [34,35]. Based on the causal constraints which restrict the set of correlations that are achievable in a given causal scenario, it is possible to define three sets: the probability distributions allowed by classical causality (the local set); those allowed by quantum mechanics (the quantum set) and those allowed by special relativity (the no-signaling set). Let us note that the no-signaling set is a superset of the quantum one, which, in turn, is a superset of the local one, see Fig. 1(b). Hence, if a particular causal model underlies an experiment and correlations that are outside the local set are registered, this unambiguously certifies the presence of nonclassical phenomena. This is the principle exploited by DI protocols, to prove the presence of quantum phenomena in a given apparatus, with no extra assumptions. A notorious example is the Bell's theorem [35], where the observation of a Bell's inequality violation, e.g., the aforementioned CHSH inequality, witnesses the presence of nonlocal correlations.

However, besides detecting nonlocal correlations, it is also important to perform optimizations over the set of quantum
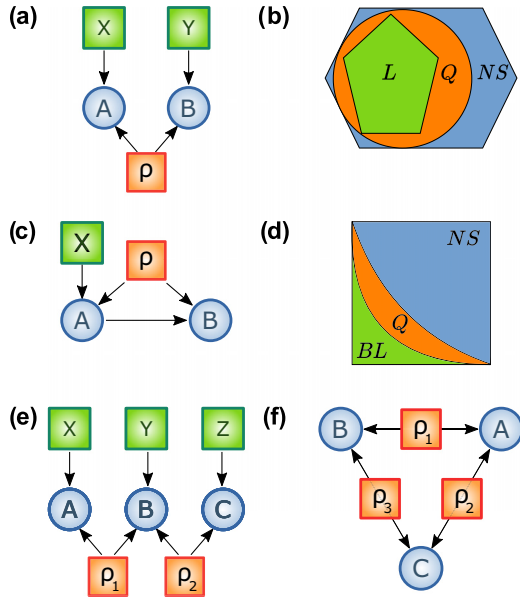
*iris.agresti@univie.ac.at

FIG. 1. Causal structures and admitted correlations set. (a) Causal structure of the standard Bell's scenario with a common source ρ, whose generated bipartite state is shared between the nodes A and B, performing measurements labeled by X and Y. (b) Sketch of the correlation (convex) sets in a standard Bell bipartite scenario [61]: local L, quantum Q, and no-signaling NS, where $L \subset Q \subset NS$. (c) Instrumental causal structure, where node A is influenced by a classical variable X (the instrument) and directly influences B. Also in this case there is only one common cause ρ. (d) Sketch of the nonconvex sets of correlations produced in the bilocality scenario. (e) DAG of the bilocality scenario, where two independent sources share correlations among three nodes: ρ_1 between A (with measurement choice X_1) and B (with measurement choice z) and ρ_2 between B and C (with measurement choice X_2). (f) The triangle scenario, where three independent sources interconnect three nodes A, B, and C, each of which performs a single measurement. This implies that, in contrast with previous scenarios, the distant parties have no measurement input.

correlations, either to find the optimal violations of Bell-like inequalities or for other applications, for example, to quantify the security of a quantum random number generator [12,24,36–43], of a quantum-key distribution protocol [32,44], or to bound the fidelity of a quantum state generated among the nodes of a quantum network [11,13,16,25,45–56]. Unfortunately, such optimization tasks are highly nontrivial, given that the quantum set is computationally hard to characterize and only approximation techniques, involving semidefinite programming (SDP), can be exploited [57]. Furthermore, such methods are limited by several restrictions. First, the objective function must be a linear function of input-output correlations. This implies that scenarios featuring more than one source of states or nonlinear entanglement witnesses [13,14,55,58,59] cannot be considered. Second, the time and memory requirements of these techniques become prohibitive as the dimension and complexity of the system grow.

In this work, we present a versatile numerical approach that circumvents these limitations. Building on the results of Ref. [60], we implement an artificial neural network–based

(ANN-based) algorithm to carry out numerical optimizations over supersets of the quantum set, see Fig. 2. At first, we benchmark our approach by considering cases well studied in the literature, involving one quantum state source, i.e., the standard Bell's scenario [61] and the instrumental scenario [17,62], see Figs. 1(a)–1(c). Then, we move a step forward and apply our technique to investigate quantum networks with independent sources, a scenario for which the standard SDP approach simply cannot be applied. At first, we consider the correlations generated in the bilocality scenario [63–65] [see Fig. 1(d)], whose DAG is depicted in Fig. 1(e) and which, despite the presence of the two independent sources, is relatively well characterized. Moving beyond, we also consider the triangle scenario with three independent sources [see Fig. 1(f)], the paradigmatic example of a causal structure whose correlations are extremely hard to characterize. Taking advantage of our method, which allows us to impose nonlinear constraints, we find additional and stronger numerical evidence supporting known results regarding the causal compatibility of quantum distributions in the triangle scenario. Furthermore, by resorting to Hoeffding's inequality [12,25,66], we show a practical application of our technique to evaluate the confidence level for the experimental violation of nonlinear causal constraints, taking into account experimental noise. Our technique can be extended to arbitrary scenarios and objective functions, thus representing a promising tool for the various uses of quantum information processing.

II. ARTIFICIAL NEURAL NETWORK OPTIMIZATION OVER QUANTUM CORRELATIONS SUPERSETS

A problem of fundamental importance at the intersection between causality and quantum information is that of distinguishing between the local, quantum and no-signaling sets of correlations allowed by a given causal network, typically represented as a directed acyclic graph (DAG) [67,68]. Even for networks with a single source of correlations, such as the Bell and the instrumental scenarios [see Figs. 1(b) and 1(c)], the characterization of such sets is a computationally demanding task and, in fact, the boundaries of the quantum set can only be approximated by a convergent hierarchy of SDPs, the Navascues-Pironio-Acín (NPA) hierarchy [69], subsumed by the positivity of a correlation matrix of increasing order. This implies that a generic optimization problem in a causal scenario involving N parties and a common state source will have the following form:

$$\text{minimize} \quad \langle C, \Gamma \rangle,$$
$$\text{subject to} \quad \langle F_k, \Gamma \rangle = b_k, \quad k = 1, \dots, d, \qquad (1)$$
$$\Gamma \succeq 0,$$

where Γ is a symmetric matrix (indicated as *moment matrix*), whose terms correspond to the expected values of products of measurement operators performed by the parties. Each product is indicated by S_i, with $i = 1, \dots, n$ and it displays a number of operators which corresponds to the level of the SDP hierarchy. A higher level implies a better approximation to the quantum set. Then, the entries $\Gamma_{i,j}$ amount to $\text{Tr}(S_i^\dagger S_j \rho)$ and they are observable, if the corresponding measurement operators commute, or nonobservable when such condition is
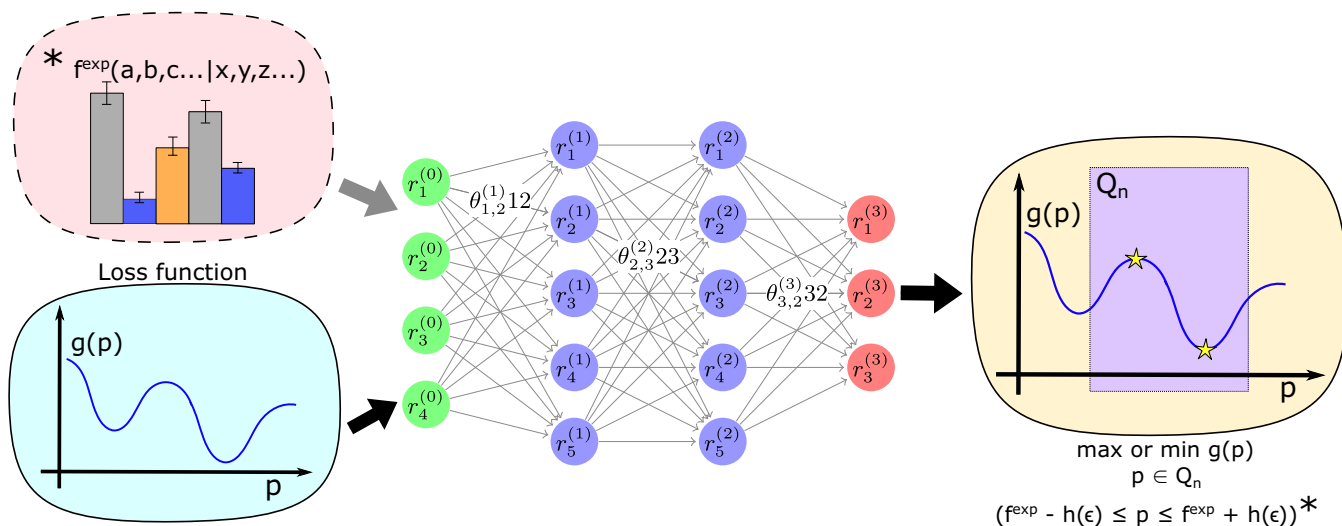
FIG. 2. ANN-based approach to the optimization of nonlinear functions over a superset of the quantum set. Our ANN optimizer works in two different configurations. In one case, it takes as input a generic nonlinear function $g(p)$ to be optimized over the superset of quantum correlations compatible with the causal scenario under analysis at the chosen level of the NPA hierarchy (see Secs. III and IV). Let us recall that, for a given causal structure, we indicate with $p(a, b, c, \ldots | x, y, z \ldots)$ or, shortly, p, the corresponding input-output probability distribution. In the second case, the ANN gets additional constraints because it is required to optimize $g(p)$ also compatibly with experimental observations, within an arbitrary confidence level 1ϵ (see Sec. V). In this instance, the ANN takes as an additional input the numerical value of the observable terms of p, which can be simulated or directly amount to experimental frequencies $f^{\text{expt}}(a, b, c, \ldots | x, y, z, \ldots)$. In both cases, the objective function $g(p)$ is set as a part of the loss function, see Eq. (3). Furthermore, the loss function embodies other constraints on the correlations terms, such as causal conditions (including no-signaling) and the confidence interval on the experimental frequencies. In the end, the ANN outputs the set of probabilities that optimize the objective function and that are compatible with the set constraints.

not satisfied, for instance, when they involve noncommuting measurements of the same party. In turn, the matrix C defines the objective function, which amounts to a linear combination of the elements of Γ. Finally, the matrices F_k and the d-dimensional vector $\mathbf{b}$ encode linear optimization constraints.

The dual problem associated with the problem in Eq. (1) can be cast as

$$\text{maximize } \langle b, y \rangle,$$
$$\text{subject to } \sum_{i=1}^{m} y_i F_i \preceq C. \qquad (2)$$

Due to strong duality [70], the problems in Eqs. (1) and (2) lead to the same solution. We reinforce that, since the aforementioned optimizations are carried out over supersets of the quantum set, the problem in Eq. (1) will only provide a lower bound of the real solution. Another drawback of this paradigmatic approach is the fact that, if we wish to consider more general networks [involving an increasing number of independent sources, such as the bilocality and triangle networks, see Figs. 1(e) and 1(f)], the problem is further complicated by the nonconvex nature of the set of correlations that no longer admit such a simple approximation. To circumvent this limitation, we combine the standard NPA approach with a machine-learning oracle that allows us to include all nonlinear constraints arising from a given causal network in the loss function of the machine optimizer.

Building on the ANN-based algorithm presented in Ref. [60], we propose an optimizer, to deal with arbitrary quantum networks and objective functions, which is directly applicable to experimental data. Our ANN architecture is that of a multilayered perceptron (MLP), which we train to maximize nonlinear functions $g(p)$ over supersets of the quantum correlations set. The essence of our approach is summarized by the conditional loss function $L(\mathbf{r}_i, \mathbf{r}_o | \theta)$. The vectors $\mathbf{r}_i$ and $\mathbf{r}_o$ represent, respectively, the state of the input and output ANN nodes, while θ indicates the set of weights and biases (which prevents the ANN to be stuck in local minima). The loss function is given by

$$\begin{cases} \Delta_{\mathcal{NS}}, & \Delta_{\mathcal{NS}} > \delta, \\ -\min \mathcal{E}, & -\min \mathcal{E} > 0, \quad \Delta_{\mathcal{NS}} < \delta, \\ -\alpha \min \mathcal{E} + \beta g(p), & -\min \mathcal{E} \leqslant 0, \quad \Delta_{\mathcal{NS}} < \delta, \end{cases} \qquad (3)$$

where $\mathcal{E}$ is the set of eigenvalues of the Γ matrix at the target level of the NPA hierarchy and with the real weights $\alpha, \beta \in [0, 1]$. Being numerically evaluated by the ANN at each iteration, the entries of Γ will be themselves functions of $(\mathbf{r}_i, \mathbf{r}_o | \theta)$. Then, $\Delta_{\mathcal{NS}}$ is a parameter used to enforce the no-signaling condition up to a precision $\delta > 0$. In detail, using the multipartite no-signaling condition, we have the following:

$$\gamma_j = \sum_{a_j} p_{A_1, \ldots, A_n | X_1, \ldots X_n}(a_1, \ldots, a_j, \ldots, a_n | x_1, \ldots, x_j, \ldots, x_n)$$
$$- \sum_{a_j} p_{A_1, \ldots, A_n | X_1, \ldots X_n}$$
$$\times (a_1, \ldots, a_j, \ldots, a_n | x_1, \ldots, x'_j, \ldots, x_n) = 0$$
$$\forall j \in [n], \quad \{a_i, \ldots, a_n\} \setminus a_j, \quad \{x_1, \ldots, x_j, x'_j, \ldots, x_n\}, \qquad (4)$$

where a_j and x_j indicate respectively the output registered by the parties and their measurement choices. Therefore, we define $\Delta_{\mathcal{NS}}$ as

$$\Delta_{\mathcal{NS}} = \sum_j |\gamma_j|, \qquad (5)$$

and $\delta = 0.001$.

The input of the ANN is then represented by the nonobservable terms of matrix Γ, while the output is the probability distribution (along with the nonobservable terms) minimizing $g(p)$ or maximizing it, if $\beta < 0$.

The working principle is that the first line of the loss in Eq. (3) constrains the probability distributions found by the ANN to be inside the no-signaling set, before solving the optimization problem. Then, the second line makes Γ as positive semidefinite as possible. When these two conditions are satisfied, implying that the correlations under scrutiny are within the demanded outer approximation of the quantum set, the third line makes the network work as an optimizer, minimizing or maximizing the desired function. Since the loss function is discontinuous, we want to avoid that, at the various steps of the training, the numerical solution that optimizes the objective function no longer defines a semidefinite positive moment matrix. We do this, on one hand, by acting on the learning rate of our algorithm, i.e., the speed with which the ANN parameters are updated and which limits the oscillations in the found solution. On the other hand, we add the term $-\alpha \min \mathcal{E}$ to maintain the condition of positive semidefiniteness while the ANN optimizes the objective function.

Let us note that the possibility of solving nonlinear optimization problems and imposing nonlinear constraints in this method comes from the fact that our ANN initializes all of the involved variables with numerical values, which are progressively changed, throughout the training, to approximate the optimal solution. Hence, the algorithm actually performs a series of feasibility problems, where all of the nonlinear constraints become linear, being numerical and, as such, they are directly enforced within the Γ matrix. Let us consider, for instance, the bilocal scenario. In this case, the causal independence nonlinear constraint $p(a, c|x, z) = p(a|x) p(c|z)$, coming from the independence between ρ_1 and ρ_2 [see Fig. 1(e)], is imposed by assigning numerical values to $p(a|x) = v_{ax}$ and $p(c|z) = v_{cz}$ and by imposing that the element of the Γ matrix corresponding to $p(a, c|x, z)$ is equal to $v_{ax} v_{cz}$. Analogous reasoning applies to nonlinear objective functions.

Hence, our approach can be defined as an instance of *self-learning*, because we feed the network with the unobservable terms of the correlation matrix Γ, all initialized randomly with a value close to zero, and it directly learns how to minimize the loss function in Eq. (3) (see Appendix A for further details).

III. SINGLE QUANTUM SOURCE SCENARIOS

To benchmark our ANN approach, we start by analyzing the following one-source scenarios: the standard bipartite Bell scenario [35] and the instrumental one [17]. We then use strong duality to certify the correct functioning of the network, analogously to Ref. [60], showing that the two solutions given by the dual and the primal problems coincide up to numerical

precision and that they provide, respectively, an approximation from above and from below of the expected result (see Appendix C).

A. Bell's scenario

In Bell-like scenarios, a common quantum source ρ generates an N-partite system, shared between N separated parties which perform local measurements on their share. In this context, we start considering the case $N = 2$, where the parties choose the measurements x and y and obtain the outcomes a and b [see Fig. 1(a)] [35]. Given this causal structure, in a classical framework (where the quantum state ρ is replaced by a classical random variable λ), the probability distribution $p(a, b|x, y)$ has testable constraints, e.g., the Clauser-Horne-Shimony-Holt (CHSH) [34] inequality:

$$\text{CHSH} = \langle A_0 B_0 \rangle + \langle A_0 B_1 \rangle + \langle A_1 B_0 \rangle - \langle A_1 B_1 \rangle \leqslant 2, \quad (6)$$

with $\langle A_x B_y \rangle \equiv \sum_{a,b} a\, b\, p(a, b|x, y)$ and which involves dichotomic measurement choices $(x, y) \in \{0, 1\}$ and outputs $(a, b) \in \{+1, -1\}$. In turn, in a quantum description, which is a more general framework, the upper bound of the constraint in Eq. (6) does not hold anymore and the upper limit is given by the Tsirelson's bound, amounting to $2\sqrt{2} \approx 2.828\,427$ [71].

We train our optimizer network by feeding it with a set of 2×10^5 nonobservable elements of the Γ matrix, at the second level of the NPA hierarchy. At the end of the training phase, we obtain the following lower and upper limits, defined by the primal and dual problem solution:

$$
\begin{aligned}
CHSH_{\text{ANN}}^{\text{primal}} &\geqslant 2.8280, \\
CHSH_{\text{ANN}}^{\text{dual}} &\leqslant 2.8285.
\end{aligned}
\qquad (7)
$$

In Fig. 3(a), we report the analytical probabilities corresponding to the analytical maximal quantum CHSH violation $[p(a, b|x, y)]$, compared with those found by the ANN $[\tilde{p}(a, b|x, y)]$. In particular, their total variational distance amounts to

$$\frac{1}{8} \sum_{\{a,b,x,y\} \in \{0,1\}} |p(a, b|x, y) - \tilde{p}(a, b|x, y)| = 0.000\,992. \quad (8)$$

In Fig. 3(b), instead, we report the maximal CHSH value predicted by the ANN (both in the dual and primal cases) as a function of the number of iterations. Let us note that, by properly choosing the learning rate and the momentum (for further details see Appendix A) it is possible to avoid being stuck in local minima and make the difference between primal and dual as small as possible.

Considering now a scenario with $N = 3$ (three parties), it is known that for dichotomic inputs and outputs there are a total of 45 nontrivial Bell's inequalities [72]. In the following, we focus on Mermin's inequality [73]

$$\mathcal{M} = \langle A_1 B_0 C_0 \rangle + \langle A_0 B_1 C_0 \rangle + \langle A_0 B_0 C_1 \rangle - \langle A_1 B_1 C_1 \rangle \leqslant 2, \quad (9)$$
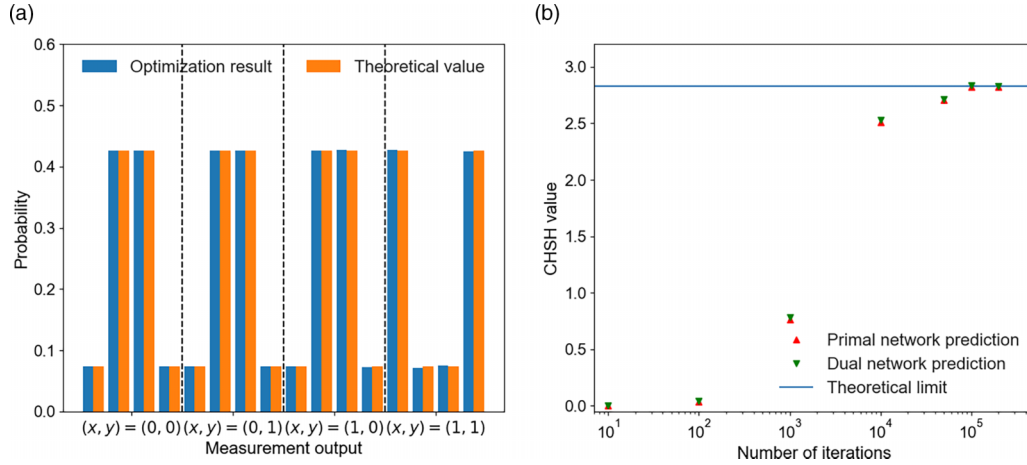
FIG. 3. ANN optimization of the CHSH inequality. In these plots, we show how the ANN-based optimizer finds the maximum violation of the CHSH inequality. (a) The blue bars indicate the correlations maximizing the CHSH inequality violation and belonging to the superset of the quantum set corresponding to the first level of the NPA hierarchy. The orange ones, instead, are the analytical values, saturating the Tsirelson's bound, i.e., $2\sqrt{2}$. On the x axis the probabilities are ordered in blocks that correspond to different choices of measurement operators. Every block contains four pairs of columns corresponding to different sets of outcomes: $(a, b) = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. (b) The red triangles represent the CHSH inequality maximal values that the ANN finds over the quantum superset corresponding to the first level of the NPA hierarchy, by solving the primal optimization problem. The green ones, instead, represent the values obtained by the ANN performing the dual problem. On the x axis, we show the number of iterations of the training. The horizontal blue line represents the analytical value corresponding to the maximum violation, i.e., the Tsirelson's bound. The fact that primal and dual solutions coincide, benchmarks the good operation of the ANN [60].

and Svetlichny's inequality [74]

$$
\begin{aligned}
\mathcal{S} = &\langle A_1 B_0 C_0 \rangle + \langle A_0 B_1 C_0 \rangle + \langle A_0 B_0 C_1 \rangle - \langle A_1 B_1 C_1 \rangle \\
&- \langle A_0 B_1 C_1 \rangle - \langle A_1 B_0 C_1 \rangle - \langle A_1 B_1 C_0 \rangle + \langle A_0 B_0 C_0 \rangle \leqslant 4.
\end{aligned}
\tag{10}
$$

With our ANN approach, still at the second level of the NPA hierarchy, through the primal optimization, we find the maximum quantum violations to be lower bounded as

$$
\begin{aligned}
\mathcal{M}_{\mathrm{ANN}}^{\mathrm{primal}} &\geqslant 3.9997, \\
\mathcal{S}_{\mathrm{ANN}}^{\mathrm{primal}} &\geqslant 5.6426,
\end{aligned}
\tag{11}
$$

an upper bounded by dual values given by

$$
\begin{aligned}
\mathcal{M}_{\mathrm{ANN}}^{\mathrm{primal}} &\leqslant 4.0006, \\
\mathcal{S}_{\mathrm{ANN}}^{\mathrm{dual}} &\leqslant 5.6639,
\end{aligned}
\tag{12}
$$

which are compatible with the known maximal quantum violations of the Mermin and Svetlichny's inequalities, amounting, respectively, to 4 and $4\sqrt{3} \approx 5.65$.

B. Instrumental scenario

Another single source scenario that we consider is the instrumental one [17,25,62,75,76], whose causal structure is shown in Fig. 1(c). As in a Bell-like causal structure, two parties share a joint quantum state. However, in this case, there is a direct causal influence from the measurement outcome a to b, the former acting as the input choice for the latter. The simplest instrumental scenario where quantum correlations fail to have a classical description happens with dichotomic measurement outcomes, but with the input X of Alice, known as the instrument [17,77], assuming three possible values

$x = 0, 1, 2$. In this case, a classical model (where the correlations are mediated by a hidden random variable λ) implies an observable constraint given by

$$
B = p(a = b|0) + p(b = 0|1) + p(a = 0, b = 1|2) \leqslant 2,
\tag{13}
$$

known as Bonet's inequality [78] and that allows for quantum violations up to $B = \frac{3+\sqrt{2}}{2} \approx 2.2071$ [17,23,62].

Training our ANN at the first level of the NPA hierarchy, we optimize the maximum violation of Bonet's inequality Eq. (13) (B_{ANN}). Analogously to the CHSH case [reported in Fig. 3(b)], also B_{ANN} increases as we increase the number of iterations reaching the following maximal values:

$$
\begin{aligned}
B_{\mathrm{ANN}}^{\mathrm{primal}} &\geqslant 2.20710, \\
B_{\mathrm{ANN}}^{\mathrm{dual}} &\leqslant 2.20720,
\end{aligned}
\tag{14}
$$

where the lower and upper limits were obtained by solving respectively the primal and dual optimization problem, for a training set size of 10^5. The total variational distance between the probability distribution obtained by the ANN optimization and the analytical distribution maximally violating Bonet's inequality amounts to $\approx 0.000\,995$.

IV. QUANTUM NETWORK SCENARIOS

Up to this point, the advantages of our machine-learning approach might not have been properly highlighted, since we have applied it to scenarios where one can use standard SDPs to maximize the objective function. In the following, we take a step further and use our ANN-based approach to study networks with independent sources, which give rise to nonconvex sets of correlations and thus require a different set of tools to be analyzed. In the classical case, algebraic

geometry methods [63–65] offer the most general approach to characterize such nonlinear constraints, but, due to their computational complexity, they become intractable already for very simple cases. Given this difficulty, a plethora of alternative approaches have been introduced to treat classical networks and some of those have also been generalized to the case of quantum ones [79–87]. However, even if able to analyze particular cases of interest, computational issues are still the main bottleneck.

A. Bilocality scenario

We start by analyzing the bilocal causal structure [63–65,82,88,89], shown in Fig. 1(d), a scenario akin to the entanglement swapping experiment where two independent sources ρ_1 and ρ_2 share subsystems among three nodes.

A classical description of this network imposes that the observed probability distribution should have a local hidden variable (LHV) model given by

$$p(a, b, c|x, y, z) = \sum_{\lambda_1, \lambda_2} p(\lambda_1)p(\lambda_2)p(a|x, \lambda_1)$$
$$\times p(b|y, \lambda_1, \lambda_2)p(c|z, \lambda_2). \quad (15)$$

Due to the independence of the sources, $p(\lambda_1, \lambda_2) = p(\lambda_1)p(\lambda_2)$, the most well-known causal constraint in this scenario amounts to a nonlinear Bell's inequality [63–65,82,88] given by

$$S = \sqrt{I_1} + \sqrt{I_2} \leqslant 1, \quad (16)$$

with

$$I_1 = \frac{1}{4} \sum_{x,z=0,1} \langle A^x C^z B^0 \rangle,$$
$$I_2 = \frac{1}{4} \sum_{x,z=0,1} (-1)^{x+z} \langle A^x C^z B^1 \rangle, \quad (17)$$

where $\langle A^x C^z B^y \rangle$ is the expected value of the measurement outcomes of the three nodes, defined as

$$\langle A^x C^z B^y \rangle = \sum_{a,c,b=0,1} (-1)^{a+c+b} p(a, c, b|x, z, y). \quad (18)$$

The maximal quantum violation of the inequality in Eq. (16) amounts to $\sqrt{2} \approx 1.4142$.

As previously mentioned, the standard NPA method cannot be applied for the bilocality scenario, due to the nonlinearity of the objective function and of the optimization constraints taking into account the independence of the sources. Remarkably, within our approach, such nonlinear constraints can be directly imposed within the loss function of the neural network solving the optimization problem as a sequence of feasibility SDP problems, as mentioned before. The value predicted by the ANN with this technique amounts to $S_{\text{ANN}} \approx 1.4134$ for a training set size of 1×10^6 (see Fig. 4).

Motivated by the ability of the ANN to recover known results we have also employed it to derive results in the bilocality scenario. For instance, we can derive upper and lower bounds for Mermin's inequality and Svetlichny's inequality
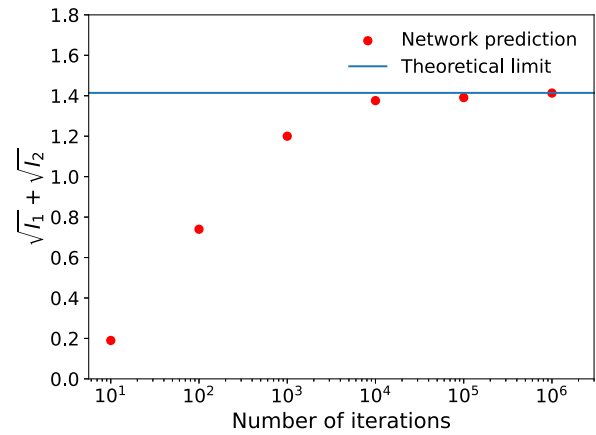

FIG. 4. ANN optimization of the bilocal inequality. Bilocal inequality maximal value predicted by the ANN as a function of the number of iterations of the training. The horizontal blue line represents the theoretical upper bound equal to $\sqrt{2} \approx 1.414\,21$.

within the bilocality scenario:

$$2.8267 \leqslant \mathcal{M} \leqslant 2.8298 \approx 2\sqrt{2},$$
$$3.9956 \leqslant \mathcal{S} \leqslant 4.0063 \approx 4, \quad (19)$$

that is, the two independent sources in the bilocality scenario cannot violate Svetlichny's inequality and, despite leading to a violation of Mermin's inequality, they cannot achieve the maximum violation $\mathcal{M} = 4$ that would be possible if a single tripartite state would be shared among the parties. As another interesting application, we use the ANN toolbox to analyze how nonlinear inequalities derived for the quantum triangle network [see Figs. 1(f) and 5(a)] perform under the bilocality constraints. For instance, the inequality

$$\langle AC \rangle + \langle BC \rangle - \langle A \rangle \langle B \rangle \leqslant 1, \quad (20)$$

is known to bound the quantum correlations in the triangle network. The ANN predicts a maximum value of 0.9999, indicating that the same inequality remains valid (with the same bound) for the bilocal causal structure.

B. Triangle scenario

Now, we consider a more complex scenario, known as *triangle quantum network*, which features three observers connected pairwise by a bipartite source, that provides a shared physical system [see Figs. 1(f) and 5(a)]. The three sources are assumed to be independent of each other, hence no tripartite information is shared between the parties. In the most general case, each observer can choose between different measurements (inputs $\{x, y, z\}$) and get an output $\{a, b, c\}$. The characterization of the set of distributions $p(a, b, c|x, y, z)$ compatible with this scenario is more complex than in the bilocality, since, due to the causal structure topology, no independence constraints can be imposed on the observed distribution. As shown in Ref. [81], this issue can be circumvented through the so-called *quantum inflation*, which allows finding upper bounds on the compatibility of a given quantum causal model with some observed correlations. In detail, we will use the cut inflation, first introduced in Ref. [81],
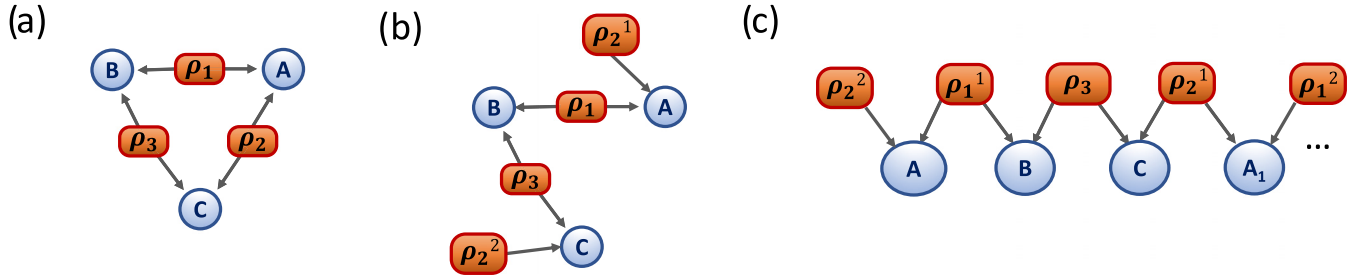
FIG. 5. Triangle scenario and cut inflation. The triangle structure is an instance of a causal structure whose compatible correlations are extremely hard to characterize since no independence conditions can be enforced. To circumvent this issue, we adopt the cut inflation causal structure, to get results that can be then translated to the triangle scenario: (a) triangle causal structure, (b) cut inflation causal structure, where the source ρ_2 is split (which then constitutes the bilocal scenario), (c) cut inflation with four observable nodes.

where we split one of the sources into two variables [see Fig. 5(b)]. Through this technique, we are able to define causal constraints, exploiting the causal independence between the peripheral parties. We can then further improve our results, by adding more copies of the original nodes in the inflated scenario, as shown in Fig. 5(c). For our results, we arrived at a maximum of nine observable nodes involved, that is third-order cut inflation, and we analyze the simplest version of the triangle scenario (with only one input and dichotomic outputs). In this framework, we study the compatibility of the so-called W distribution, defined as follows:

$$p_W(a, b, c) = \begin{cases} \frac{1}{3}, & a+b+c = 1 \\ 0, & a+b+c \neq 1. \end{cases} \quad (21)$$

with the quantum triangle scenario. Indeed, it was proven that this distribution is not realizable both in the classical [81] and quantum triangle scenarios [83]. In further detail, when mixing the W distribution with white noise, the resulting statistics

$$P_{W,v}(a, b, c) = vP_W(a, b, c) + (1-v)/8 \quad (22)$$

does not have a quantum realization for visibilities higher than $3(2 - \sqrt{3}) \simeq 0.8039$. Note that this is a numerical bound, which is not known to be tight. Hence, we use our ANN approach to test it, using different levels of the NPA hierarchy and orders of the cut inflation (see Appendix B) and enforcing stricter constraints, with respect to those which can be imposed through standard SDP optimizations involving scalar extension [90].

Adopting our method, for the second-order cut inflation (i.e., with eight observable nodes) and at the fourth level of the hierarchy we obtain $v \simeq 0.8043$, which is compatible with the previous result. To improve this result, we solved the problem associated with a third-order cut inflation at the fifth level of the NPA hierarchy and with nine observable nodes (see Fig. 5), finding again a very similar bound amounting to

$$v \simeq 0.8033. \quad (23)$$

This indicates that the cut inflation cannot improve over known results, since increasing the inflation and NPA levels does not lead to any significant changes in the known bounds, far off what could be argued to be beyond any numerical approximations. We notice that, for this last result, only the primal problem has been solved, due to memory issues, related to the high number of optimization constraints. However,

as previously mentioned, the precision of the primal solution is highly dependent on the learning rate of the network and can be made negligible. Moreover, we can give an estimate of this precision, by looking at the difference between primal and dual results of the same problem, but related to lower levels of the NPA hierarchy and lower orders of inflation. In particular, for six nodes and third level of the NPA hierarchy, we find that the difference between the results is lower than 10^{-4}.

Furthermore, note that for the scenario with no inputs, the no-signaling condition [see Eq. (4)] does not need to be enforced. Thus we use a reduced form of the conditional loss in Eq. (3), given by

$$\begin{cases} -\min \mathcal{E}, & -\min \mathcal{E} > 0, \\ -\alpha \min \mathcal{E} + \beta g(p), & -\min \mathcal{E} \leqslant 0. \end{cases} \quad (24)$$

V. APPLICATION TO EXPERIMENTAL DATA

After showing the applicability of our approach in a variety of scenarios involving single or multiple quantum sources, from a theoretical point of view, we prove next that our method can also be used as a certification tool directly applicable to experimental data. In particular, given a causal scenario whose experimental statistics violate given testable constraints, our aim is to find the minimum violation which is, at the same time, compatible with experimental data and with quantum mechanics. As opposed to usual error propagation, where one assumes that the experimental data is distributed according to a specific distribution, e.g., the Poissonian one, here, we only require our data to be independent and identically distributed (the *iid* assumption). Thus, we estimate the experimental errors resorting to Hoeffding's inequality [66], which provides an upper bound on the probability that *iid* random variables deviate from their expected value more than a given threshold. Using this inequality and following the analysis provided in Ref. [26], for each experimental frequency f_j^{expt}, obtained from n_j registered counts corresponding to a given input and output configuration j, we can say, with a confidence level $1 - \epsilon$, that

$$f_j^{\text{expt}} - h_j(\epsilon) \leqslant p_j \leqslant f_j^{\text{expt}} + h_j(\epsilon), \quad (25)$$

where p_j is the expected value and $h_j(\epsilon)$ amounts to

$$h_j(\epsilon) = \sqrt{\frac{-\ln\epsilon}{2n_j}}, \quad (26)$$

Considering constraints like the one in Eq. (25) for every j, we find the minimum quantum violation which complies with them and that belongs to the superset of quantum correlations corresponding to the considered NPA level. This allows us to achieve a lower bound on the experimental quantum violation, with the only *iid* assumption.

To do that, we modify the ANN structure, adding a number of outputs corresponding to the probability terms. Those outputs are activated using a hyperbolic tangent loss function and thus constrained in the interval $[-1, 1]$ and used as weights w_i to modulate the value of the input behavior inside the Hoeffding's error. Hence, the loss function will be the same as in Eq. (3) and, for example, in the ANN used for the bilocality scenario, the 64 outputs will represent the weights w_i for the Hoeffding's errors h_i. We also modify the inputs of the network, that here are constituted by the experimental statistics. Let us note that, in this case, we need to train the network to minimize a general probability distribution compatible with the considered causal structure. Thus we numerically generate input samples emulating the experiment causal structure (see Appendix A for further details) and inject them into our network. The Hoeffding's errors [i.e., Eq. (26)] are instead given as hyperparameters to the network. Then, using the loss function with $g(p) = S$ [see Eq. (16)] and the Γ corresponding to the correlation matrix associated with the second level of the NPA hierarchy, we train the ANN on the following probabilities:

$$p'_i = |p_i + w_i h_i| \qquad (27)$$

where $i = (a, b, c|x, y, z)$, which we normalize in the end.

Hence, the ANN learns to minimize the quantum bilocal inequality violation over the quantum set, taking into account Hoeffding's error. Let us note that these lower bounds could be made fully DI by using another statistical error given by the Azuma's inequality [91]. Unfortunately, this inequality is more sensitive to noise, so we could only obtain trivial bounds in our case. To collect the experimental statistics, we physically implemented Bell's and the bilocality scenarios. To do this, we resorted to the versatile photonic platform described in Refs. [20,26] and shown in Fig. 6. The main advantage of this platform resides in the fact that it can be easily adapted to implement arbitrary networks, just by slightly changing its topology (for more details, see Appendix D). Then, starting from the statistics collected in Bell's scenario, we also recovered the instrumental correlations, through a particular procedure of postselection. Indeed, the instrumental scenario can be mapped to a standard Bell's scenario [23], through the following mapping for $x = (0, 1)$:

$$p_{\text{instr}}(a, b|x) = p_{\text{Bell}}(a, b|x, y = a) \qquad (28)$$

and $p_{\text{instr}}(a, b|x = 2) = \delta_{a,0} p_{\text{Bell}}(b|y)$.

For each of the two employed entangled sources, we obtained the following experimental values for Bell's and the instrumental violations:

$$
\begin{aligned}
CHSH_1 &= 2.656\,20 \pm 0.009\,72 > 2, \\
CHSH_2 &= 2.580 \pm 0.010 > 2, \\
B_1 &= 2.1640 \pm 0.0024 > 2, \\
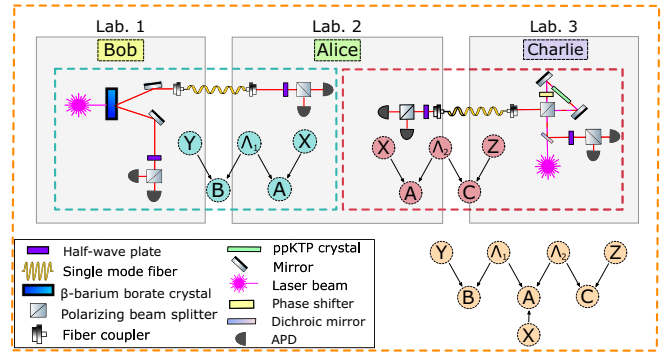B_2 &= 2.1451 \pm 0.0025 > 2.
\end{aligned} \qquad (29)
$$

FIG. 6. Experimental apparatus. We employ a platform already introduced in Refs. [20,26] to implement the bilocality scenario. In detail, we employ three different laboratories. Those held respectively by Bob and Charlie (Laboratories 1 and 3) are equipped with one quantum state source and one measurement station, while the one held by Alice (Laboratory 2) only includes two measurement stations. The source in Laboratory 1, Λ_1, sends one photon to Bob's measurement station, while source Λ_2, in Laboratory 3, sends one to Charlie. Moreover, both sources send the other photon to Alice, namely to Laboratory 2, through a $\approx$30 m long single-mode fiber. The two employed sources are spontaneous parametric down-conversion sources implemented through a beta barium borate type II crystal (Laboratory 1) and through a periodically poled potassium tytanil phosphate in a Sagnac loop (Laboratory 3). The measurement stations are composed by a half-wave plate, followed by a polarizing beam splitter, which allows performing any projective measurement, in the polarization degree of freedom, of the form $\cos(\alpha)\sigma_x + \sin(\alpha)\sigma_z$, where σ_x and σ_z are Pauli matrices. Then, to implement two instances of the Bell's scenario, we separately used two of the three laboratories at a time, i.e., only one state source and two measurement stations.

The uncertainties reported in Eq. (29) are evaluated in a device-dependent way, assuming that the underlying statistics is Poissonian. To drop this assumption and estimate the confidence level of the nonclassicality within our experiment device-independently, we feed the ANN with the experimental statistics collected in the three aforementioned scenarios. Thus we obtain a neural network prediction of the minimum possible violation of target Bell and Bell-like inequalities compatible with the experimental causal structure.

For both cases, the nonclassicality holds up to a confidence level of $1 - \epsilon$, with $\epsilon = 10^{-25}$. Then, for the bilocality scenario, whose experimental value amounts to 1.3080 ± 0.0035, we were able to certify the nonclassicality of the experimental statistics up to a confidence level of $1 - 10^{-17}$. We report the obtained results in Fig. 7, where we show the comparison between the probability distributions generated by the network and the experimental one, as well as the predicted values, as a function of the number of iterations.

VI. DISCUSSION

In this work, we devised a method based on artificial neural networks to optimize arbitrary nonlinear functions over supersets of quantum correlations. Building on the oracle introduced in Ref. [60], we propose a general framework allowing the analysis of notably difficult causal structures.
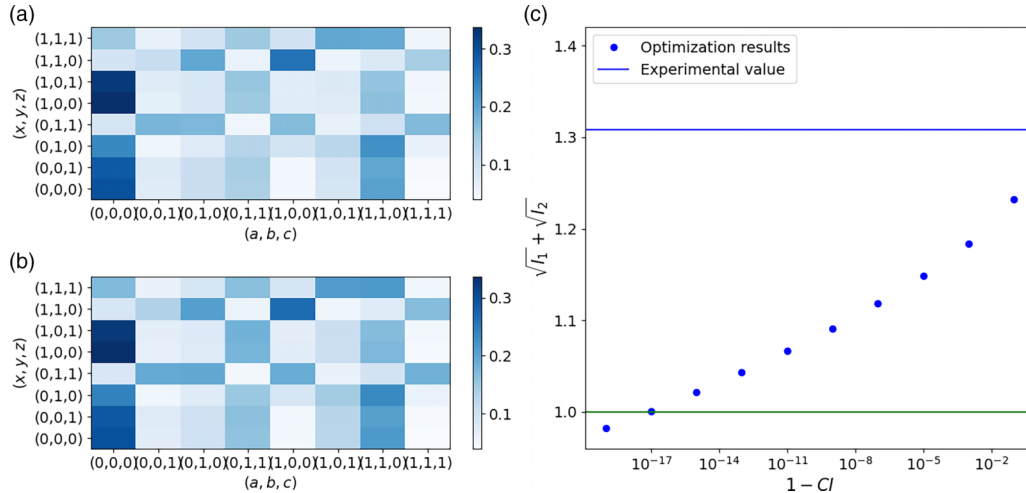
FIG. 7. Experimental results in the bilocality scenario. Comparison between experimental statistics $f^{\text{expt}}(a, b, c | x, y, z)$ (a) and values predicted by the ANN using a confidence interval $CI = 1 - 10^{-6}$. (b) The columns and rows represent, respectively, (a, b, c) and (x, y, z) triples. (c) Lower bounds on the bilocal inequality violation predicted by the ANN tested on the experimental values as a function of the complementary of the confidence interval $\epsilon = 1 - CI$. The green line represents the minimum value needed for the network to certify nonlocality. The blue line indicates the experimental violation value. Due to the training with numerically simulated experimental data, the ANN can then be employed to obtain lower bounds for any experimental violation obtained in the bilocal scenario, i.e. with no additional time requirements. At the same time, this implies that $\Delta_{\mathcal{NS}}$ [see Eq. (3)] will be of the same order of magnitude of the Hoeffding's uncertainties.

Nicely, it can be directly applied to data, finding confidence levels on experimental quantum violations of causal constraints, taking into account real noisy cases, as we practically demonstrated by realizing different experiments employing a versatile photonic setup.

We benchmark our framework by first applying it to a number of known relevant cases, including the standard Bell, the instrumental, the bilocal and the triangle causal structures. We then go beyond previously known results, and prove, for instance, that two independent sources of quantum states cannot violate Svetlichny's inequality, paradigmatic in the study of genuine multipartite nonlocality. We also enhanced the confidence in previously known numerical results for the triangle case, by considering higher levels on the quantum inflations and on the NPA hierarchy. It is worth mentioning that even though here we have focused on the quantum case, our technique can also be adapted to analyze classical causal networks, basically by imposing commutativity on the operators in the moment matrix. Our algorithm thus offers a complementary machine-learning framework to those that have been recently introduced to analyze the classical case [60,84,92–95].

In summary, the remarkable features of our method are, first of all, its versatility: it can indeed be applied to any quantum network, featuring independent sources, i.e., with nonlinear objective functions and optimization constraints. Second, when applied to experimental statistics, this approach does not require any data regularization, as the network can be fed with raw experimental data. Third, in terms of time requirements, our method proves to be more efficient than standard SDP methods, since the ANN time demands (which, for linear optimization problems are comparable to those of SDP-based methods) are limited to the training process. Hence, once trained, the network can be used multiple times with no further time demands. Due to these advantages, this algorithm paves the way for different certification protocols for quantum-based technologies. In particular, it can prove particularly apt for communication and cryptography tasks and can be straightforwardly extended to larger and more complex quantum networks.

APPENDIX A: NEURAL NETWORK ARCHITECTURE

Here, we give a more detailed description of the artificial neural network (ANN) architecture presented in this work, consisting of a multilayer perceptron (MLP), as well as the parameters used for the implementation of the proposed method.

A MLP is a feedforward artificial neural network that consists of a series of layers that are directly connected to each other in a noncyclic manner and whose total number defines the depth d of the ANN. The number of neurons present in each layer l defines its width w_l and all neurons in the same layer are activated through a function σ_l. Hence, the output of layer l is given by

$$r^{(l+1)} = \sigma_{l+1}(W^{(l+1)}r^{(l)} + b^l), \qquad \text{(A1)}$$

TABLE I. ANN architecture and parameters (primal quantum networks). The width refers to the number of neurons of the hidden layers, the input and output layer dimension is defined by the NPA hierarchy level for each scenario. α is the momentum and η_0 the initial learning rate.

Scenario	NPA level	Width	Length	α	η_0
Bell	2	300	8	0.8	10^{-4}
Instrumental	2	500	8	0.8	10^{-4}
Bilocality	2	560	8	0.8	10^{-4}
Tripartite Bell	2	560	8	0.9	10^{-4}

TABLE II. ANN architecture and parameters (dual quantum networks). The width refers to the number of neurons of the hidden layers, the input and output layer dimension is defined by the NPA hierarchy level for each scenario. α is the momentum and η_0 the initial learning rate.

Scenario	NPA level	Width	Length	α	η_0
Bell	2	300	8	0.8	10^{-4}
Instrumental	2	1000	8	0.8	10^{-4}
Tripartite Bell	2	1000	8	0.8	10^{-4}

where $r^{(l)}$ indicates the neurons of layer l and W^{l+1} represents the matrix containing the weights between layer l and $l + 1$. Finally, b_l stands for the bias vector, which is an additional set of weights that has the effect of shifting the activation function of target neurons by a constant amount. The bias vector is needed since the network layers perform linear operations on the inputs. Hence, if the input is null, the network could only produce a zero output. Furthermore, these additional learning parameters increase the overall accuracy.

In the following, the trainable parameters of the model will be collectively defined as weights and denoted by $\theta = \{W^l, b^l\}$. The action of finding the appropriate weights that minimize a function of the training data $R = \{r_i^{(0)}\}$, known as loss function $\mathcal{L}(R|\theta)$, is called *training* and there are many algorithms available to perform this task. The most commonly used one is the *stochastic gradient descent*, in which a random subset of the training data, called batch R_j, is used to update the weights as

$$\theta' = \theta - \eta \nabla_\theta \mathcal{L}(R_j|\theta), \qquad (A2)$$

where η represents the learning rate. In this work, we use a variant of such a method, the *online gradient descent*, where each batch is made up of a single element. Let us note that, if trained on larger batches, the model could overlook the presence of the discontinuities leading to a negative gradient in batches where the objective function f is not minimized correctly. We also apply a momentum to the updates, which is an additive term that considers the updates at the previous step $\Delta \theta$ with weight α, thus using a rule given by

$$\theta' = \theta - \eta \nabla_\theta \mathcal{L}(r_j|\theta) + \alpha \Delta \theta. \qquad (A3)$$

We choose the online gradient descent algorithm since the loss $L(r_1, r_d|\theta)$ that we use [reported also in Eq. (3)] is the following step function:

$$\begin{cases} \Delta_{\mathcal{NS}}, & \Delta_{\mathcal{NS}} > \epsilon, \\ -\min \mathcal{E}, & -\min \mathcal{E} > 0, \quad \Delta_{\mathcal{NS}} < \epsilon, \\ -\alpha \min \mathcal{E} + \beta f(p), & -\min \mathcal{E} \leqslant 0, \quad \Delta_{\mathcal{NS}} < \epsilon, \end{cases} \qquad (A4)$$

where $\mathcal{E}$ stands for the eigenvalues of $\Gamma(\mathbf{r}_i, \mathbf{r}_o|\theta)$, which corresponds to the Γ matrix at the target level of the NPA hierarchy and with the real weights $\alpha, \beta \in [0, 1]$. In turn, $\Delta_{\mathcal{NS}}$ is a parameter used to enforce the no-signaling condition up to a precision $\epsilon > 0$.

To increase the precision of the method, we also add a learning rate decay of 0.2 per 2×10^4 rounds in all the single source networks. Detailed information on the ANN architecture is provided in Tables I, II, and III, concerning the

application of our method to theoretical bounds on quantum networks, and Table IV, for the ANNs used on experimental data. In all our architectures, the momentum is fixed. We use exponential linear unit activation functions for all the ANNs hidden layers.

In the ANNs used for the function optimizations, the outputs are composed of a series of neurons activated through a normalized exponential function, or softmax, corresponding to the predicted behavior, and a number of neurons equal to the inputs. Those elements act like errors for the inputs, thus the activation function used for them is the hyperbolic tangent. The nonphysical elements of the moment matrix, γ_i^{NP}, can then be evaluated as

$$\gamma_i^{NP} = r_i^{(0)} + r_{m-1}^{(n)}, \qquad (A5)$$

where m is the total number of neurons of the output layer. The idea behind the introduction of that new output is that, with this new degree of freedom, the network is able to learn the full structure of the problem, without needing external information to find the maximum of the Bell inequality under scrutiny.

The most natural way to make the ANN learn by itself with such a structure is to have all the input samples equal to the zero vector. In all the scenarios that involve nonlinear objective functions, however, this method tends to lead the network into local minima in the early phases of the training. We thus use zero vectors as inputs to obtain upper bounds for Eqs. (6), (9), (10), and (13) and random inputs extracted in the $[-10^{-3}, 10^{-3}]$ interval to optimize Eq. (16). The training set size used is 2×10^5 for all the single source networks and 1×10^6 for the networks involving more sources. In general, it

TABLE III. ANN architecture and parameters (primal quantum networks, triangle). The width refers to the number of neurons of the hidden layers, the input and output layer dimension is defined by the NPA hierarchy level for each scenario. α is the momentum and η_0 the initial learning rate. In this scenario it was possible to build a dual network only for the six node network due to memory reasons. The six-node NPA level three dual network has width equal to 4000, $\alpha = 0.8$, and $\eta_0 = 1 \times 10^{-4}$.

Nodes	NPA level	Width	Length	α	η_0
6	3	300	8	0.8	10^{-4}
8	4	1000	8	0.8	10^{-4}
9	5	1000	8	0.8	10^{-4}
9	6	10000	8	0.8	10^{-4}

TABLE IV. ANN architecture, experimental data. The width refers to the number of neurons of the hidden layers, the input and output layer dimension is defined by the NPA hierarchy level for each scenario. As usual α is the momentum and η the learning rate.

Scenario	NPA level	Width	Length	α	η
Bell	1	80	8	0.8	10^{-4}
Instrumental	1	120	8	0.8	10^{-4}
Bilocality	2	560	8	0.8	10^{-4}

is advisable to reduce the interval from which the samples are extracted and increasing the number of training samples as the number of nonphysical elements of the moment matrix grows. This was the method used in the optimizations of causal constraints in the quantum network scenarios considered.

In turn, the ANNs applied to the experimental data are slightly different. The first set of output neurons of those ANNs represents the missing elements of the moment matrix, activated through rectified linear units. The second part of the outputs acts as error on the probability distribution and, as in the other family of ANNs, it is activated through hyperbolic tangent units. The error is given as an unchanging hyperparameter to the ANN. To generate the training samples we simulated the bilocality scenario, implemented through the experimental apparatus shown in Fig. 6, considering that, in the ideal case, the peripheral parties implement the following projective measurements:

$$\left|\Psi_x^0\right\rangle = \cos(x\pi/4)|0\rangle + \sin(x\pi/4)|1\rangle,$$
$$\left|\Psi_x^1\right\rangle = \cos(x\pi/4)|1\rangle - \sin(x\pi/4)|0\rangle, \tag{A6}$$

for each setting $x_i = x$, and the one performed in the central node:

$$\left|\Phi_y^0\right\rangle = \cos\frac{(2y+1)\pi}{8}|0\rangle + \sin\frac{(2y+1)\pi}{8}|1\rangle,$$
$$\left|\Phi_y^0\right\rangle = \cos\frac{(2y+1)\pi}{8}|1\rangle - \sin\frac{(2y+1)\pi}{8}|0\rangle, \tag{A7}$$

for each setting $y_i = y$. Then, we consider that the state generated in Laboratory 1 and Laboratory 3 is given by

$$|\psi\rangle = |\psi^-\rangle_1 \otimes |\psi^-\rangle_2, \tag{A8}$$

where $|\psi^-\rangle = \frac{|01\rangle - |01\rangle}{\sqrt{2}}$. To simulate the presence of imperfections in the experimental apparatus, white noise was added to each generated state, through the following model:

$$\rho = \rho_1 \otimes \rho_2, \tag{A9}$$

with

$$\rho_1 = v|\psi^-\rangle\langle\psi^-| + (1-v)\frac{\mathbb{I}}{2}. \tag{A10}$$

On top of that, to emulate the statistic of the experiment the behavior were extracted inside the Hoeffding's uncertainties [see Eq. (25)].

In Figs. 8 and 9, we report the quantum violation lower bound, respectively, in the Bell and instrumental scenario, compatible with the experimental statistics and with quantum mechanics, versus the confidence level Cl given by Hoeffding's inequality [66], more precisely versus $1 - Cl$. These
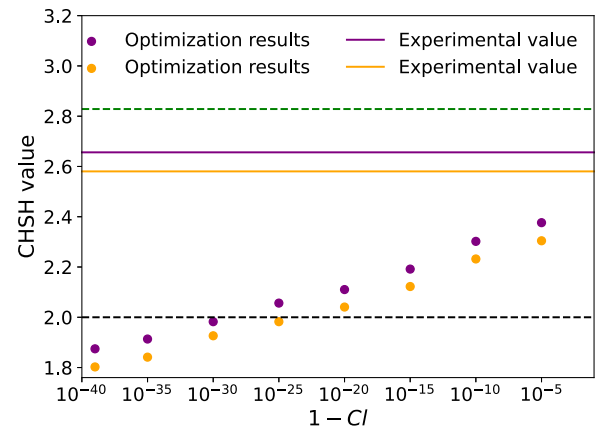

FIG. 8. CHSH experimental violation confidence (Hoeffding inequality). In this plot, we report the lower bound on the experimental violation of the CHSH inequality, obtained considering different confidence levels in the result, through Hoeffding's inequality. This inequality requires the *iid* assumption, i.e., that the runs are independent and identically distributed. The green dashed line stands for the theoretical upper limit, admitted by quantum mechanics. The orange and purple solid lines represent the experimental violations obtained through the apparatus depicted in Fig. 6, considering the two sources separately (i.e., two instances of the Bell scenario).

results are valid under the so-called *iid* assumption, i.e., the experimental runs have to be independent and identically distributed, so, strictly speaking, nonfully device-independent.

To have full device-independence, a different inequality must be adopted, i.e., Azuma's inequality [91]. Unfortunately, due to the lower robustness to noise of this inequality, only trivial lower bounds could be extracted from our experimental statistics. Hence, in Figs. 10 and 11, we report nontrivial lower
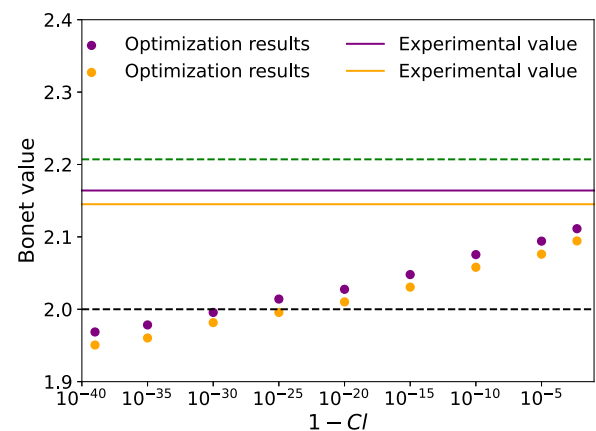

FIG. 9. Bonet experimental violation confidence (Hoeffding inequality). In this plot, we report the lower bound on the experimental violation of the Bonet inequality, obtained considering different confidence levels in the result, through Hoeffding's inequality. This inequality requires the *iid* assumption, i.e., that the runs are independent and identically distributed. The green dashed line stands for the theoretical upper limit, admitted by quantum mechanics. The orange and purple solid lines represent the experimental violations obtained through the apparatus depicted in Fig. 6, considering the two sources separately (i.e., two instances of Bell scenario, lifted to obtain instrumental correlations [62]).
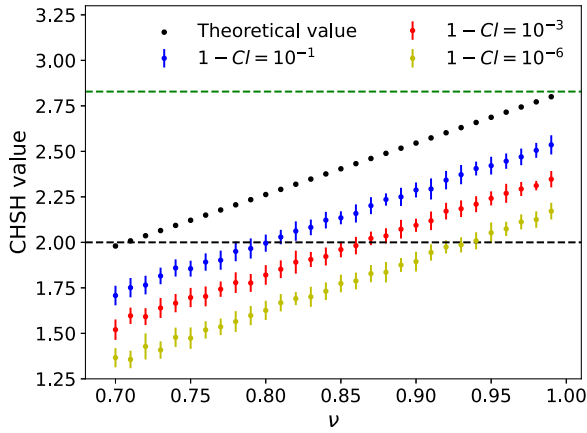
FIG. 10. CHSH experimental violation confidence (Azuma inequality). In this plot, we report the lower bound on the experimental violation of the CHSH inequality, obtained considering different confidence levels in the result, through Azuma's inequality, and different levels of white noise [see Eq. (A10)]. This inequality drops the *iid* assumption, i.e., that the runs are independent and identically distributed, and hence these results are fully device-independent. The green dashed line stands for the theoretical upper limit, admitted by quantum mechanics. The black one, instead, stands for the theoretical upper limit, admitted by classical physics. The data used in this case are numerically simulated.

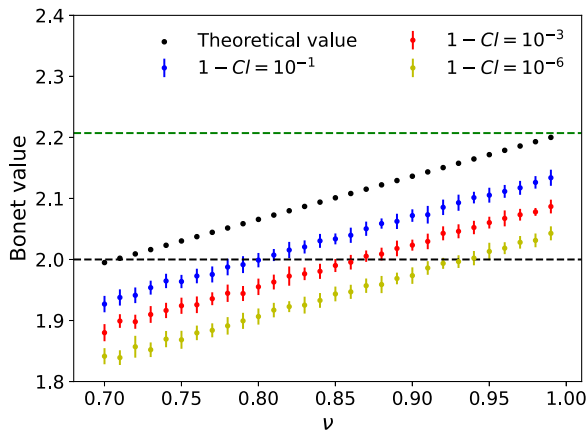bounds evaluated through Azuma's inequality on numerically simulated noisy data.

FIG. 11. Bonet experimental violation confidence (Azuma inequality.) In this plot, we report the lower bound on the experimental violation of the Bonet inequality, obtained considering different confidence levels in the result, through Azuma's inequality, and different levels of white noise [see Eq. (A10)]. This inequality drops the *iid* assumption, i.e., that the runs are independent and identically distributed, and hence these results are fully device-independent. The green dashed line stands for the theoretical upper limit, admitted by quantum mechanics. The black one, instead, stands for the theoretical upper limit, admitted by classical physics. The data used in this case are numerically simulated.

APPENDIX B: AUTOMATED CREATION OF THE MOMENT MATRIX Γ

In this section, we present the algorithm used to generate the moment matrix and the matrices used to encode the optimization constraints used in the considered dual problems. This algorithm builds a dictionary that associates the expected values of the operators present in the density matrix with the elements of the density matrix where they appear.

We define the unitary operators with which we build the moment matrix as alphanumerical strings of the form "X_n", where the letter defines the party involved and the number subscript refers to the measurement setting. Given the number of parties, inputs and the level of the NPA hierarchy, the monomial set can be evaluated as the combination of a number of operators equal to the NPA level. By doing this, we obtain a monomial set, which is larger than the true one, since elements like "$A_0B_0A_1$" and "$A_0A_1B_0$" are both considered, so we need to remove one of them, as well as elements like "A_0A_0", which would correspond to 1. Those tasks can be done by imposing the commutativity of different operators and the fact that the operators are unitary. We thus remove all elements with subsequent repetition.

Let us recall that the elements of the density matrix can be obtained as

$$\Gamma_{i,j} = \text{Tr}(S_i^\dagger S_j \rho), \quad i, j = 1, \dots, n, \quad \text{(B1)}$$

where S_i represents the ith operator in the monomial set and n is the dimension of the density matrix Γ. Since our operators are Hermitian, i.e., $S = S^\dagger$, in order to define the complex conjugate of the operator products, we revert the monomial set elements, so that, for instance, "A_0A_1" becomes "A_1A_0." Then, after removing the term repetitions, we enforce the commutativity and unitarity constraints, by putting some terms equal to others, e.g., $A_0B_0 = B_0A_0$. In the end, we generate a dictionary, whose keys represent the operators, and the values indicate the positions, where those operators appear within the Γ matrix.

As previously stated, the strength of our approach stands in the possibility of enforcing the causal independence nonlinear constraints. The aforementioned dictionary is hence divided into three parts. First, given the independent parties as inputs, the algorithm inserts a "*" character between all the independent operators inside the dictionary keys. Then all remaining keys are checked to find possible nonmeasurable elements. That is all elements that contain any alphabetical subsequent repetition, with different subscripts, for example "$A_0A_1B_0$."

Thus, in the end, we get three dictionaries: (i) the one associated with the measurable elements (ii) the nonmeasurable, and (iii) the elements with causal independence constraints. Then, the dictionaries are imported and used as a tool to initialize variables inside the density matrix. Here we define a function "expected_value" to evaluate the expected value of the physical elements, while the nonphysical elements are always associated with variables of the ANN. For what concerns the causally independent terms we use the additional character "*" to split the string, then we obtain the value of the element under exam as a product of the values associated with each string. We do that using another custom function "causal_indep_prod." To give some examples:

(1) Given the physical element $\langle A_0 \rangle$ the expected value is evaluated, then this value is multiplied for the measurable dictionary element "A_0."

(2) Given the nonphysical element $\langle A_0 A_1 \rangle$ one of the optimization variables of the network is multiplied for the nonmeasurable dictionary element "$A_0 A_1$."

(3) Given two causally independent peripheral parties A and C, the element $\langle A_0 C_0 \rangle$ is correctly viewed as $\langle A_0 \rangle \langle C_0 \rangle$, so the expected values $\langle A_0 \rangle$ and $\langle C_0 \rangle$ are evaluated and then multiplied by each other. The result is multiplied by the causal independence dictionary element "$A_0 C_0$."

The final step is to consider the possibility of an inflated scenario. The user chooses the number of nodes, by creating a list of ordered alphabetical characters, both for the inflated and noninflated nodes. All the monomials that have to be accounted for as variables, being associated with noninjectable terms, are those that include multiple copies of the same source. That is, defining n the number of originals nodes, the monomials that contain subsequent therms equal to n or sequential characters distant to each other n or $n-1$. All the noninjectable monomials are then added to the nonphysical element dictionary and removed from the physical element one. Finally, we substitute the copies with the original alphabet in all the dictionaries except for the monomials associated with noninjectable terms. To give some examples, in the triangle cut inflation scenario with six nodes "A,""B,""C,""D,""E,""F":

(1) Given the physical injectable element $\langle D_0 \rangle$ this value appears in the physical element dictionary as "A_0."

(2) Given the noninjectables $\langle A_0 B_0 C_0 \rangle$ and $\langle D_0 E_0 F_0 \rangle$ both elements are added to the nonphysical element dictionary. Notice that there are two distinct variables associated with the dictionary keys "$A_0 B_0 C_0$" and "$D_0 E_0 F_0$."

Note that the causal independence was already enforced previously and that some noninjectable elements in the linear constraint scenario become injectable measurable elements. For example $\langle A_0 C_0 \rangle = \langle A_0 \rangle \langle C_0 \rangle$ is now injectable being a product of injectable terms.

The code is available at Ref. [96].

APPENDIX C: IMPLEMENTATION OF THE DUAL PROBLEM OPTIMIZATION

Given the primal formulation of a standard SDP problem, given by

$$
\begin{aligned}
& \text{minimise} \langle C, \Gamma \rangle, \\
& \text{subject to} \langle F_k, \Gamma \rangle = b_k, \quad k = 1, \ldots, N, \\
& \qquad \Gamma \succeq 0,
\end{aligned}
\tag{C1}
$$

the dual formulation reads as follows:

$$
\begin{aligned}
& \text{maximise} \langle b, y \rangle, \\
& \text{subject to} \sum_{i=1}^{m} y_i F_i \preceq C.
\end{aligned}
\tag{C2}
$$

To implement the dual problem with our artificial neural network, we use the same dictionaries defined for the primal, to define the matrix $C - \sum_{i=1}^{m} y_i F_i$, which is then constrained to be semidefinite positive.

Let us note that, for the primal formulation, the F_k matrices did not need to be defined, since all the constraints were directly enforced on the elements of matrix Γ, i.e., on the corresponding dictionary elements. This is not the case for the dual problem, where, for each F_i matrix, a unique variable y_i has to be initialized for each primal problem constraints. So, for example, let us consider the dictionary element corresponding to the key "A_0":

$$
A_0 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},
\tag{C3}
$$

where the 1 values indicate all the positions where the element $\langle A_0 \rangle$ can be found in the Γ matrix. In the dual formulation of the problem, we need to impose that all the nonzero matrix elements of "A_0" are equal to each other. We do that by means of equalities of the form $\langle F_i, \Gamma \rangle = 0$. Hence, for example, the constraints matrices F_i that enforce the equality constraints in Eq. (C3) will be

$$
F_1 = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}, \tag{C4}
$$

such that

$$
\begin{aligned}
\langle F_1, \Gamma \rangle = \Gamma_{1,2} - \Gamma_{1,3} = 0 &\rightarrow v_{1,2} = v_{1,3}, \\
\langle F_2, \Gamma \rangle = \Gamma_{1,2} - \Gamma_{2,3} = 0 &\rightarrow v_{1,2} = v_{2,3}.
\end{aligned}
\tag{C5}
$$

Notably, as in the primal case, we are able to impose nonlinear constraints since, at each step of the training, the neural network algorithm acts as a numerical feasibility problem. As such, any nonlinear constraint in the optimization problem is reduced to a fixed numerical value. To give an example the causal independence constraint $\langle A_0 C_0 \rangle = \langle A_0 \rangle \langle C_0 \rangle$ can be directly enforced since, at the start of the training, the network uses a prediction of the probability distribution to evaluate the loss function.

APPENDIX D: DETAILS OF EXPERIMENTAL APPARATUS

Our experimental apparatus, depicted in Fig. 6, employs three separated laboratories linked through two $\approx$30-m-long single-mode fibers. Two of them are equipped with one source of polarization-entangled photon pairs and one measurement station. The third laboratory, instead, is equipped with two measurement stations.

The source in Laboratory 1 employs type-II spontaneous parametric down-conversion (SPDC) to generate pairs of polarization-entangled photons with a wavelength equal to $\lambda = 785$ nm, through a β barium borate crystal which is pumped, in a pulsed regime, by a $\lambda = 392.5$ nm laser beam. Instead, in Laboratory 2, we have a periodically poled potassium titanyl phosphate crystal, pumped in a continuous-wave regime by a $\lambda = 404$ nm laser, which generates polarization-entangled photon pairs at $\lambda = 808$ nm. Both sources are

optimized to generate a two-qubit maximally entangled state, e.g., the singlet state $|\psi^-\rangle$, where the computational basis ($|0\rangle$ and $|1\rangle$) is encoded in the horizontal and vertical photon polarization states ($|H\rangle$ and $|V\rangle$), hence $|\psi^-\rangle = \frac{|HV\rangle - |VH\rangle}{\sqrt{2}}$.

The events recorded at each measurement station are analyzed through a central time-tagger and software that coordinates the counters located in the different laboratories and recognizes the counts occurring within a given time window as coincidence events of distant detectors. In detail, the optimal window for two-fold coincidences is 1.05 ns, while the one for four-fold coincidences, namely, events where two two-fold coincidences are recorded within the chosen time window, is 1.033 μs. We now go into further details of the implementation of the bilocality scenario, since the Bell scenario is just a simpler instance of this configuration. Laboratories 1 and 3, equipped with one quantum state source (respectively, ρ_1 and ρ_2) and one measurement station each, are held respectively by Bob and Charlie. Laboratory 2, which is equipped with two measurement stations, is held by Alice. Each measurement station is made up by a half-wave plate (HWP) and a polarizing beam splitter (PBS), allowing one to perform polarization projective measurements of the

form $\cos(4\theta)\sigma_z + \sin(4\theta)\sigma_x$, where σ_x and σ_z are the standard Pauli operators, by simply rotating the HWP of an angle θ with respect to its optical axis. Single photons are revealed through avalanche photodiode detectors. The three laboratories are connected through two $\approx$30-m-long single-mode fibers, as shown in Fig. 6. Both laboratories deliver a photon to the other one and perform projective measurements on the photon they kept and on the one they received. In detail, the observers' measurement operators are the following: $A_0 = A_0^1 \otimes A_0^2 = \sigma_z \otimes \sigma_z$ and $A_1 = A_1^1 \otimes A_1^2 = \sigma_x \otimes \sigma_x$; $B_0 = -\frac{(\sigma_x + \sigma_z)}{\sqrt{2}}$ and $B_1 = \frac{(\sigma_x - \sigma_z)}{\sqrt{2}}$; $C_0 = -\frac{(\sigma_x + \sigma_z)}{\sqrt{2}}$ and $C_1 = \frac{(\sigma_x - \sigma_z)}{\sqrt{2}}$, where the superscripts indicate the source-generating the subsystem. Such operators are, up to unitary transformations, those maximizing the CHSH inequality violation. On the other hand, we use source ρ_1 and source ρ_2 separately to implement two different instances of a Bell scenario [Fig. 1(a)]. From the same statistics, we can also retrieve instrumental correlations [see the corresponding DAG in Fig. 1(c)], by lifting the Bell scenario, as in Refs. [23,62] [see Eq. (28)].

Then, by considering the three laboratories together, we implement the bilocality scenario [see the corresponding DAG in Fig. 1(e)].

---

[1] N. Gisin and R. Thew, Quantum communication, Nat. Photonics **1**, 165 (2007).

[2] M. A. Nielsen and I. Chuang, *Quantum Computation and Quantum Information* (American Association of Physics Teachers, 2002).

[3] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Quantum cryptography, Rev. Mod. Phys. **74**, 145 (2002).

[4] A. K. Ekert, Quantum Cryptography Based on Bell's Theorem, Phys. Rev. Lett. **67**, 661 (1991).

[5] T. Jennewein, C. Simon, G. Weihs, H. Weinfurter, and A. Zeilinger, Quantum Cryptography with Entangled Photons, Phys. Rev. Lett. **84**, 4729 (2000).

[6] C. Schimpf, M. Reindl, D. Huber, B. Lehner, S. F. Covre da Silva, M. Vyvlecka, P. Walther, and A. Rastelli, Quantum cryptography with highly entangled photons from semiconductor quantum dots, Sci. Adv. **7**, eabe8905 (2021).

[7] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, The security of practical quantum key distribution, Rev. Mod. Phys. **81**, 1301 (2009).

[8] F. Xu, X. Ma, Q. Zhang, H.-K. Lo, and J.-W. Pan, Secure quantum key distribution with realistic devices, Rev. Mod. Phys. **92**, 025002 (2020).

[9] J. Barrett, L. Hardy, and A. Kent, No Signaling and Quantum Key Distribution, Phys. Rev. Lett. **95**, 010503 (2005).

[10] A. Acín, N. Gisin, and L. Masanes, From Bell's Theorem to Secure Quantum Key Distribution, Phys. Rev. Lett. **97**, 120405 (2006).

[11] I. Šupić and J. Bowles, Self-testing of quantum systems: A review, Quantum **4**, 337 (2020).

[12] S. Pironio, A. Acín, S. Massar, A. B. de La Giroday, D. N. Matsukevich, P. Maunz, S. Olmschenk, D. Hayes, L. Luo, and T. A. Manning *et al.*, Random numbers certified by Bell's theorem, Nature (London) **464**, 1021 (2010).

[13] G. Carvacho, F. Andreoli, L. Santodonato, M. Bentivegna, R. Chaves, and F. Sciarrino, Experimental violation of local causality in a quantum network, Nat. Commun. **8**, 14775 (2017).

[14] D. J. Saunders, A. J. Bennet, C. Branciard, and G. J. Pryde, Experimental demonstration of nonbilocal quantum correlations, Sci. Adv. **3**, e1602743 (2017).

[15] M. Ringbauer, C. Giarmatzi, R. Chaves, F. Costa, A. G. White, and A. Fedrizzi, Experimental test of nonlocal causality, Sci. Adv. **2**, e1600162 (2016).

[16] F. Andreoli, G. Carvacho, L. Santodonato, M. Bentivegna, R. Chaves, and F. Sciarrino, Experimental bilocality violation without shared reference frames, Phys. Rev. A **95**, 062315 (2017).

[17] R. Chaves, G. Carvacho, I. Agresti, V. Di Giulio, L. Aolita, S. Giacomini, and F. Sciarrino, Quantum violation of an instrumental test, Nat. Phys. **14**, 291 (2018).

[18] E. Polino, I. Agresti, D. Poderini, G. Carvacho, G. Milani, G. B. Lemos, R. Chaves, and F. Sciarrino, Device-independent test of a delayed choice experiment, Phys. Rev. A **100**, 022111 (2019).

[19] Q.-C. Sun, Y.-F. Jiang, B. Bai, W. Zhang, H. Li, X. Jiang, J. Zhang, L. You, X. Chen, and Z. Wang *et al.*, Experimental demonstration of non-bilocality with truly independent sources and strict locality constraints, Nat. Photonics **13**, 687 (2019).

[20] D. Poderini, I. Agresti, G. Marchese, E. Polino, T. Giordani, A. Suprano, M. Valeri, G. Milani, N. Spagnolo, and G. Carvacho *et al.*, Experimental violation of *n*-locality in a star quantum network, Nat. Commun. **11**, 2467 (2020).

[21] A. Suprano, D. Poderini, E. Polino, I. Agresti, G. Carvacho, A. Canabarro, E. Wolfe, R. Chaves, and F. Sciarrino, Experimental genuine tripartite nonlocality in a quantum triangle network, PRX Quantum **3**, 030342 (2022).

[22] G. Carvacho, E. Roccia, M. Valeri, F. B. Basset, D. Poderini, C. Pardo, E. Polino, L. Carosini, M. B. Rota, J. Neuwirth,

S. F. C. da Silva, A. Rastelli, N. Spagnolo, R. Chaves, R. Trotta, and F. Sciarrino, Quantum violation of local causality in an urban network using hybrid photonic technologies, Optica **9**, 572 (2022).

[23] I. Agresti, G. Carvacho, D. Poderini, L. Aolita, R. Chaves, and F. Sciarrino, Experimental connection between the instrumental and Bell inequalities, Proceedings **12**, 27 (2019).

[24] I. Agresti, D. Poderini, L. Guerini, M. Mancusi, G. Carvacho, L. Aolita, D. Cavalcanti, R. Chaves, and F. Sciarrino, Experimental device-independent certified randomness generation with an instrumental causal structure, Commun. Phys. **3**, 110 (2020).

[25] I. Agresti, D. Poderini, B. Polacchi, N. Miklin, M. Gachechiladze, A. Suprano, E. Polino, G. Milani, G. Carvacho, and R. Chaves *et al.*, Experimental test of quantum causal influences, Sci. Adv. **8**, eabm1515 (2022).

[26] I. Agresti, B. Polacchi, D. Poderini, E. Polino, A. Suprano, I. Šupić, J. Bowles, G. Carvacho, D. Cavalcanti, and F. Sciarrino, Experimental robust self-testing of the state generated by a quantum network, PRX Quantum **2**, 020346 (2021).

[27] M.-C. Chen, C. Wang, F.-M. Liu, J.-W. Wang, C. Ying, Z.-X. Shang, Y. Wu, M. Gong, H. Deng, and F.-T. Liang *et al.*, Ruling Out Real-Valued Standard Formalism of Quantum Theory, Phys. Rev. Lett. **128**, 040403 (2022).

[28] Z.-D. Li, Y.-L. Mao, M. Weilenmann, A. Tavakoli, H. Chen, L. Feng, S.-J. Yang, M.-O. Renou, D. Trillo, and T. P. Le *et al.*, Testing Real Quantum Theory in an Optical Quantum Network, Phys. Rev. Lett. **128**, 040402 (2022).

[29] E. Polino, D. Poderini, G. Rodari, I. Agresti, A. Suprano, G. Carvacho, E. Wolfe, A. Canabarro, G. Moreno, and G. Milani *et al.*, Experimental nonclassicality in a causal network without assuming freedom of choice, Nat. Commun. **14**, 909 (2023).

[30] N.-N. Wang, A. Pozas-Kerstjens, C. Zhang, B.-H. Liu, Y.-F. Huang, C.-F. Li, G.-C. Guo, N. Gisin, and A. Tavakoli, Certification of non-classicality in all links of a photonic star network without assuming quantum mechanics, arXiv:2212.09765.

[31] S. Pironio, V. Scarani, and T. Vidick, Focus on device independent quantum information, New J. Phys. **18**, 100202 (2016).

[32] A. Acín, N. Brunner, N. Gisin, S. Massar, S. Pironio, and V. Scarani, Device-Independent Security of Quantum Cryptography against Collective Attacks, Phys. Rev. Lett. **98**, 230501 (2007).

[33] J.-D. Bancal, M. Navascués, V. Scarani, T. Vértesi, and T. H. Yang, Physical characterization of quantum devices from nonlocal correlations, Phys. Rev. A **91**, 022115 (2015).

[34] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Proposed Experiment to Test Local Hidden-Variable Theories, Phys. Rev. Lett. **23**, 880 (1969).

[35] J. S. Bell, On the Einstein Podolsky Rosen paradox, Phys. Phys. Fiz. **1**, 195 (1964).

[36] P. Bierhorst, E. Knill, S. Glancy, Y. Zhang, A. Mink, S. Jordan, A. Rommal, Y.-K. Liu, B. Christensen, and S. W. Nam *et al.*, Experimentally generated randomness certified by the impossibility of superluminal signals, Nature (London) **556**, 223 (2018).

[37] L. Shen, J. Lee, J.-D. Bancal, A. Cerè, A. Lamas-Linares, A. Lita, T. Gerrits, S. W. Nam, V. Scarani, and C. Kurtsiefer *et al.*, Randomness Extraction from Bell Violation with Continuous Parametric Down-Conversion, Phys. Rev. Lett. **121**, 150402 (2018).

[38] U. Vazirani and T. Vidick, Fully device independent quantum key distribution, Commun. ACM **62**, 133 (2019).

[39] V. Scarani, The device-independent outlook on quantum physics, Acta Phys. Slovaca **62**, 347 (2012).

[40] R. Ramanathan, F. G. Brandão, K. Horodecki, M. Horodecki, P. Horodecki, and H. Wojewódka, Randomness Amplification under Minimal Fundamental Assumptions on the Devices, Phys. Rev. Lett. **117**, 230501 (2016).

[41] U. Vazirani and T. Vidick, Certifiable quantum dice: or, true random number generation secure against quantum adversaries,' in *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing* (2012), pp. 61–76.

[42] R. Arnon-Friedman and R. Renner, de Finetti reductions for correlations, J. Math. Phys. **56**, 052203 (2015).

[43] J.-D. Bancal, L. Sheridan, and V. Scarani, More randomness from the same data, New J. Phys. **16**, 033011 (2014).

[44] R. Arnon-Friedman, F. Dupuis, O. Fawzi, R. Renner, and T. Vidick, Practical device-independent quantum cryptography via entropy accumulation, Nat. Commun. **9**, 459 (2018).

[45] D. Poderini, E. Polino, G. Rodari, A. Suprano, R. Chaves, and F. Sciarrino, *Ab initio* experimental violation of Bell inequalities, Phys. Rev. Res. **4**, 013159 (2022).

[46] I. Šupić, D. Cavalcanti, and J. Bowles, Device-independent certification of tensor products of quantum states using single-copy self-testing protocols, Quantum **5**, 418 (2021).

[47] D. Mayers and A. Yao, Self testing quantum apparatus, Quantum Inf. Comput. **4**, 273 (2004).

[48] I. Šupić, R. Augusiak, A. Salavrakos, and A. Acín, Self-testing protocols based on the chained Bell inequalities, New J. Phys. **18**, 035013 (2016).

[49] I. Šupić, A. Coladangelo, R. Augusiak, and A. Acín, Self-testing multipartite entangled states through projections onto two systems, New J. Phys. **20**, 083041 (2018).

[50] M. O. Renou, J. Kaniewski, and N. Brunner, Self-Testing Entangled Measurements in Quantum Networks, Phys. Rev. Lett. **121**, 250507 (2018).

[51] T. H. Yang, T. Vértesi, J.-D. Bancal, V. Scarani, and M. Navascués, Robust and Versatile Black-Box Certification of Quantum Devices, Phys. Rev. Lett. **113**, 040401 (2014).

[52] J. Kaniewski, Analytic and Nearly Optimal Self-Testing Bounds for the Clauser-Horne-Shimony-Holt and Mermin Inequalities, Phys. Rev. Lett. **117**, 070402 (2016).

[53] J. Kaniewski, Self-testing of binary observables based on commutation, Phys. Rev. A **95**, 062323 (2017).

[54] I. Šupić, J. Bowles, M.-O. Renou, A. Acín, and M. J. Hoban, Quantum networks self-test all entangled states, Nat. Phys., 1 (2023).

[55] C. M. Lee and M. J. Hoban, Towards Device-Independent Information Processing on General Quantum Networks, Phys. Rev. Lett. **120**, 020504 (2018).

[56] F. Andreoli, G. Carvacho, L. Santodonato, R. Chaves, and F. Sciarrino, Maximal qubit violation of n-locality inequalities in a star-shaped quantum network, New J. Phys. **19**, 113020 (2017).

[57] M. Navascués, S. Pironio, and A. Acín, Bounding the Set of Quantum Correlations, Phys. Rev. Lett. **98**, 010401 (2007).

[58] K. Sen, C. Srivastava, and U. Sen, Measurement-device-independent nonlinear entanglement witnesses, arXiv:2106.05796.

[59] O. Gühne and N. Lütkenhaus, Nonlinear Entanglement Witnesses, Phys. Rev. Lett. **96**, 170502 (2006).

[60] T. Kriváchy, Y. Cai, J. Bowles, D. Cavalcanti, and N. Brunner, High-speed batch processing of semidefinite programs with feedforward neural networks, New J. Phys. **23**, 103034 (2021).

[61] N. Brunner, D. Cavalcanti, S. Pironio, V. Scarani, and S. Wehner, Bell nonlocality, Rev. Mod. Phys. **86**, 419 (2014).

[62] T. Van Himbeeck, J. B. Brask, S. Pironio, R. Ramanathan, A. B. Sainz, and E. Wolfe, Quantum violations in the instrumental scenario and their relations to the Bell scenario, Quantum **3**, 186 (2019).

[63] A. Tavakoli, P. Skrzypczyk, D. Cavalcanti, and A. Acín, Nonlocal correlations in the star-network configuration, Phys. Rev. A **90**, 062109 (2014).

[64] C. Branciard, N. Gisin, and S. Pironio, Characterizing the Nonlocal Correlations Created via Entanglement Swapping, Phys. Rev. Lett. **104**, 170401 (2010).

[65] C. Branciard, D. Rosset, N. Gisin, and S. Pironio, Bilocal versus nonbilocal correlations in entanglement-swapping experiments, Phys. Rev. A **85**, 032119 (2012).

[66] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. **58**, 13 (1963).

[67] R. Chaves, R. Kueng, J. B. Brask, and D. Gross, Unifying Framework for Relaxations of the Causal Assumptions in Bell's Theorem, Phys. Rev. Lett. **114**, 140403 (2015).

[68] M. S. Leifer and R. W. Spekkens, Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference, Phys. Rev. A **88**, 052130 (2013).

[69] M. Navascués, S. Pironio, and A. Acín, A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations, New J. Phys. **10**, 073013 (2008).

[70] L. Vandenberghe and S. Boyd, Semidefinite programming, SIAM Rev. **38**, 49 (1996).

[71] B. S. Cirel'son, Quantum generalizations of Bell's inequality, Lett. Math. Phys. **4**, 93 (1980).

[72] C. Śliwa, Symmetries of the Bell correlation inequalities, Phys. Lett. A **317**, 165 (2003).

[73] N. D. Mermin, Extreme Quantum Entanglement in a Superposition of Macroscopically Distinct States, Phys. Rev. Lett. **65**, 1838 (1990).

[74] G. Svetlichny, Distinguishing three-body from two-body nonseparability by a Bell-type inequality, Phys. Rev. D: Part. Fields **35**, 3066 (1987).

[75] R. Nery, M. Taddei, R. Chaves, and L. Aolita, Quantum Steering Beyond Instrumental Causal Networks, Phys. Rev. Lett. **120**, 140408 (2018).

[76] M. Gachechiladze, N. Miklin, and R. Chaves, Quantifying Causal Influences in the Presence of a Quantum Common Cause, Phys. Rev. Lett. **125**, 230401 (2020).

[77] J. Pearl, On the testability of causal models with latent and instrumental variables, in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (1995), pp. 435–443, https://dl.acm.org/doi/abs/10.5555/2074158.2074208.

[78] B. Bonet, Instrumentality tests revisited, in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (2001), pp. 48–55, https://dl.acm.org/doi/10.5555/2074022.2074029.

[79] A. Tavakoli, A. Pozas-Kerstjens, and M.-O. Renou *et al.*, Bell nonlocality in networks, Rep. Prog. Phys. **85**, 056001 (2021).

[80] R. Chaves, C. Majenz, and D. Gross, Information–theoretic implications of quantum causal structures, Nat. Commun. **6**, 5766 (2015).

[81] E. Wolfe, R. W. Spekkens, and T. Fritz, The inflation technique for causal inference with latent variables, J. Causal Inference **7**, 20170020 (2019).

[82] R. Chaves, Polynomial Bell Inequalities, Phys. Rev. Lett. **116**, 010402 (2016).

[83] E. Wolfe, A. Pozas-Kerstjens, M. Grinberg, D. Rosset, A. Acín, and M. Navascués, Quantum Inflation: A General Approach to Quantum Causal Compatibility, Phys. Rev. X **11**, 021043 (2021).

[84] T. Kriváchy, Y. Cai, D. Cavalcanti, A. Tavakoli, N. Gisin, and N. Brunner, A neural network oracle for quantum nonlocality problems in networks, npj Quantum Inf. **6**, 70 (2020).

[85] M.-O. Renou, E. Bäumer, S. Boreiri, N. Brunner, N. Gisin, and S. Beigi, Genuine Quantum Nonlocality in the Triangle Network, Phys. Rev. Lett. **123**, 140401 (2019).

[86] J. Åberg, R. Nery, C. Duarte, and R. Chaves, Semidefinite Tests for Quantum Network Topologies, Phys. Rev. Lett. **125**, 110505 (2020).

[87] R. Chaves, G. Moreno, E. Polino, D. Poderini, I. Agresti, A. Suprano, M. R. Barros, G. Carvacho, E. Wolfe, and A. Canabarro *et al.*, Causal networks and freedom of choice in Bell's theorem, PRX Quantum **2**, 040323 (2021).

[88] D. Rosset, C. Branciard, T. J. Barnea, G. Pütz, N. Brunner, and N. Gisin, Nonlinear Bell Inequalities Tailored for Quantum Networks, Phys. Rev. Lett. **116**, 010403 (2016).

[89] N. Gisin, Q. Mei, A. Tavakoli, M. O. Renou, and N. Brunner, All entangled pure quantum states violate the bilocality inequality, Phys. Rev. A **96**, 020304(R) (2017).

[90] A. Pozas-Kerstjens, R. Rabelo, Ł. Rudnicki, R. Chaves, D. Cavalcanti, M. Navascués, and A. Acín, Bounding the Sets of Classical and Quantum Correlations in Networks, Phys. Rev. Lett. **123**, 140503 (2019).

[91] K. Azuma, Weighted sums of certain dependent random variables, Tohoku Math. J. Second Ser. **19**, 357 (1967).

[92] D.-L. Deng, Machine Learning Detection of Bell Nonlocality in Quantum Many-Body Systems, Phys. Rev. Lett. **120**, 240402 (2018).

[93] A. Canabarro, S. Brito, and R. Chaves, Machine Learning Nonlocal Correlations, Phys. Rev. Lett. **122**, 200401 (2019).

[94] K. Bharti, T. Haug, V. Vedral, and L.-C. Kwek, Machine learning meets quantum foundations: A brief survey, AVS Quantum Sci. **2**, 034101 (2020).

[95] A. Pozas-Kerstjens, N. Gisin, and M.-O. Renou, Proofs of Network Quantum Nonlocality Aided by Machine Learning, Phys. Rev. Lett. **130**, 090201 (2023).

[96] https://github.com/Nicoljno/Machine-learning-based-device-independent-certification-of-quantum-networks.