Quantum Science and Technology

PAPER • **OPEN ACCESS**

Validating multi-photon quantum interference with finite data

To cite this article: Fulvio Flamini *et al* 2020 *Quantum Sci. Technol.* **5** 045005

View the article online for updates and enhancements.

IOP ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 141.108.6.18 on 20/01/2021 at 10:59

Quantum Science and Technology

OPEN ACCESS

RECEIVED
7 April 2020

REVISED
16 June 2020

ACCEPTED FOR PUBLICATION
25 June 2020

PUBLISHED
16 July 2020

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Validating multi-photon quantum interference with finite data

Fulvio Flamini[1,2,8], **Mattia Walschaers**[3], **Nicolò Spagnolo**[1], **Nathan Wiebe**[4,5,6], **Andreas Buchleitner**[7] **and Fabio Sciarrino**[1]

[1] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 5, I-00185 Roma, Italy
[2] Institut für Theoretische Physik, Universität Innsbruck, Technikerstraße 21a, 6020 Innsbruck, Austria
[3] Laboratoire Kastler Brossel, Sorbonne Université, CNRS, Collège de France, ENS-PSL Research University, 4 Place Jussieu, 75005 Paris, France
[4] Microsoft Research, Station Q Quantum Architectures and Computation Group, Redmond, WA, United States of America
[5] Pacific Northwest National Laboratory, Richland, WA, United States of America
[6] Department of Physics, University of Washington, Seattle, WA, United States of America
[7] Physikalisches Institut, Albert-Ludwigs-Universität Freiburg, Hermann-Herder-Strasse 3, 79104 Freiburg, Germany
[8] Author to whom any correspondence should be addressed.

E-mail: fulvio.flamini@uibk.ac.at

Keywords: Boson Sampling, multi-photon interference, validation, machine learning, quantum technologies

Abstract

Multi-particle interference is a key resource for quantum information processing, as exemplified by Boson Sampling. Hence, given its fragile nature, an essential desideratum is a solid and reliable framework for its validation. However, while several protocols have been introduced to this end, the approach is still fragmented and fails to build a big picture for future developments. In this work, we propose an operational approach to validation that encompasses and strengthens the state of the art for these protocols. To this end, we consider the Bayesian hypothesis testing and the statistical benchmark as most favorable protocols for small- and large-scale applications, respectively. We numerically investigate their operation with finite sample size, extending previous tests to larger dimensions, and against two adversarial algorithms for classical simulation: the mean-field sampler and the metropolized independent sampler. To evidence the actual need for refined validation techniques, we show how the assessment of numerically simulated data depends on the available sample size, as well as on the internal hyper-parameters and other practically relevant constraints. Our analyses provide general insights into the challenge of validation, and can inspire the design of algorithms with a measurable quantum advantage.

1. Introduction

A quantum computational advantage occurs when a quantum device starts outperforming its best classical counterpart on a given specialized task [1, 2]. Intermediate models [3–6] and platforms [7–12] have been proposed to achieve this regime, largely reducing the physical resources required by universal computation. The technological race towards quantum computational advantage goes nonetheless hand-in-hand with the development of classical protocols capable to discern genuine quantum information processing [13–17]. The intertwined evolution of these two aspects has been highlighted in particular by Boson Sampling [3, 18], where several protocols have been introduced [19–35] and experimentally tested [31–46] to rule out non-quantum processes. Boson Sampling, in its original formulation [3], consists in sampling from a probability distribution that can be related to the evolution of indistinguishable photons in a linear-optical interferometer. Recent analyses suggested reasonable thresholds in the number of photons n to surpass classical algorithms [47, 48, 50].

While the sampling task itself has been thoroughly analyzed in computational complexity theory, we still lack a comparable understanding when it comes to its validation. However, it is clear from a practical perspective that any computational problem designed to demonstrate quantum advantage needs to be formulated together with a set of validation protocols which account for the physical ramifications and

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

resources required for its implementation. For instance, while small-scale examples can be validated by direct solution of the Schrödinger equation and using statistical measures such as cross-entropy [6], this is prohibitively expensive to debug a faulty Boson sampler. Moreover, for Boson Sampling a deterministic certification is impossible [24] by the very definition of the problem [20]. Hence, it is crucial to develop debugging tools, as well as tests to exclude undesired hypotheses on the system producing the output, that are computationally affordable and experimentally feasible. Furthermore, due to random fluctuations inherent to any finite-size problem, a validation cannot be considered reliable until sufficient physical resources are spent to obtain reasonable experimental uncertainties. Ultimately, no computational problem can provide evidence of quantum advantage unless quantitative validation criteria can be stated.

In this work, we investigate the problem of validating multi-photon quantum interference in realistic scenarios with finite data. The paper is structured as follows: first, we discuss possible ambiguities in the validation of Boson Sampling, which play a crucial role in large-size experiments. Then, building upon state-of-the-art validation protocols, we address the above considerations with a more quantitative analysis. We describe a practical approach to validation that makes the most of the limited physical resources available. Specifically, we study the use of the statistical benchmark [30] and the Bayesian hypothesis testing [31] to validate n-photon interference for large and small n, respectively. We numerically investigate their operation against classical algorithms to simulate quantum interference, with a particular focus on the number of measurements. The reported analysis strengthens the need for a well-defined approach to validation, both to demonstrate quantum advantage and to assist applications that involve multi-photon states.

2. Validation of Boson Sampling: framework

Our aim, in the context of Boson Sampling, consists in the unambiguous identification of a quantum advantage in a realistic scenario. We focus on the task of *validation*, or verification, whose aim is to check if measured experimental data is compatible with what can be expected from a given physical model. Validation generally requires fewer resources and is, thus, more appropriate for practical applications than full certification, which is exponentially hard in n for Boson Sampling [20, 51]. In both cases, these claims must follow a well-defined protocol to distill experimental evidence that is accepted by the community under jointly agreed criteria [52] (figure 1). As we discuss below and in section 3, we propose an application-oriented approach to validation that takes into consideration the limited physical resources, be them related to the evaluation of permanents [53] or to finite sample size [51]. In fact, without such well-defined approaches, obstacles or ambiguities may arise in large-scale experiments, as we highlight in the following. For instance, not all validation protocols are computationally efficient, which is a strong limitation for future multi-photon applications or high-rate real-time monitoring. Also, a theoretically scalable validation protocol may still be experimentally impractical due to large instrumental overheads or large prefactors that enter the scaling law.

Given two validation protocols $\mathcal{V}_1$ and $\mathcal{V}_2$ to rule out the same physical hypothesis or model, which conclusion can be drawn if they agree for a data set of given size and unexpectedly disagree when we add more data? In principle we can accept or reject a data set when we reach a certain level of confidence, but which action is to be taken if this threshold is not reached after a large number of measurement events (which hereafter we refer to as the 'sample size')? Shall we proceed until we pass that level, shall we reject it or shall we make a guess on the available data? Finally, what if the classical algorithm becomes more effective in simulating Boson Sampling for larger data sets, as for Markov chains [47], or for longer processing times, as for adversarial machine learning algorithms [55] that could exploit specific vulnerabilities of validation protocols?

However artificial some of the above questions may seem, such skeptical approach was indeed already adopted [25] and addressed [26–30, 35–37] with the mean-field sampler (see appendix A): all these considerations are necessary to strengthen the claim of quantum advantage. Under the above premise, we therefore identify the following crucial features to be assessed in any decision on acceptance or rejection:

(a) Sample size $\mathcal{S}$. The strength of a validation protocol is affected by the limited number $\mathcal{S}$ of collected events, as compared to the total number of distinct n-photon output events. While this limitation is not relevant for small-scale implementations, due to (i) the then low dimension of Hilbert space, (ii) a high level of control and (iii) reduced losses, it represents one of the main bottlenecks for the actually targeted large-scale instances [56]. It is thus desirable to assess the robustness and the resilience of a protocol under such incomplete-sampling effects, to quantify the impact of always strictly finite experimental resources on the actual applicability range of the protocol. We therefore propose to define a (minimal) threshold sample size $\mathcal{S}$ which must be available for validation. Given a set of $\mathcal{S}$ events, a validation protocol must be capable to give a reliable answer within a certain confidence level.

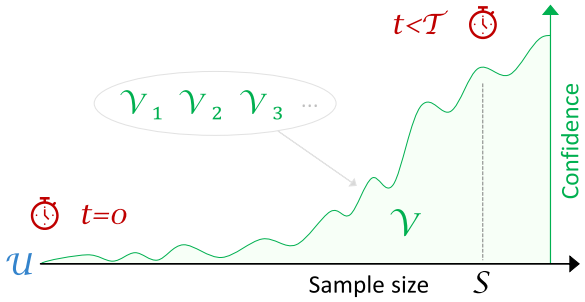

Figure 1. To demonstrate quantum advantage, reliable and realistic approaches to validation need to be defined. Boson Sampling should be validated with well-defined sampling time ($\mathcal{T}$) and sample size ($\mathcal{S}$), since the efficacy of validation protocols ($\mathcal{V}$) changes with the number of measured events after the unitary evolution ($\mathcal{U}$).

(b) Available sampling time $\mathcal{T}$. While the sampling rate is nearly constant for current quantum and classical approaches [48], *de facto* making the time $\mathcal{T}$ not relevant, it cannot be excluded that future algorithms may process data and output all events at once. The very quality of the simulation, i.e. the similarity to quantum Boson Sampling in a given metric, could also improve with processing time [47, 55]. Ultimately, $\mathcal{T}$ must be treated as an independent parameter with respect to $\mathcal{S}$, while at the same time it should be adapted to the sample size required for a reliable validation.

(c) Unitary $\mathcal{U}$. Unitary evolutions should be drawn Haar-randomly by a third agent, at the start of the competition to avoid any preprocessing. This agent, the *validator* ($\mathcal{V}$), uses specific validation protocols to decide whether a sample is compatible with quantum operation.

In the thus defined setting, a data set is said validated according to the following rule (figure 1(a)):

Boson Sampling is validated if, collecting $\mathcal{S}$ events in time $\mathcal{T}$ from some random unitary $\mathcal{U}$, it is accepted by all selected validators $\mathcal{V}$.

Given a unitary and a set of validation protocols, we are then left with the choice of $\mathcal{S}$ and $\mathcal{T}$, which need be plausible for technological standards. Demanding to sample $\mathcal{S}$ events in time $\mathcal{T}$, these thresholds in fact limit the size of the problem (n, m) for an experimental implementation. As for the time $\mathcal{T}$, one possibility, feasible for quantum experiments, could be for instance one hour. Within this time, a quantum device will probably output events at a nearly constant rate, while a classical computer can output them at any rate allowed by its clock cycle time. The choice of the sample size $\mathcal{S}$ is instead more intricate, since a value too high collides with the limited $\mathcal{T}$, while a value too low implies an unreliable validation $\mathcal{V}$. With these or further considerations [57], classical and quantum samplers should agree upon a combination of $(n, m, \mathcal{S}, \mathcal{T})$ that allows them to validate their operation.

3. Validation with finite sample size

In this section, we investigate a convenient approach to validation that distinguishes between two regimes: until $n \sim 30$ (section 3.1) and from $n \sim 30$ (section 3.2). In each section, we will first summarize the main ideas behind their operation. Then, we will discuss their performance for various (n, m), highlighting strengths and limitations, by numerically simulating experiments with finite sample size and distinguishable or indistinguishable photons.

3.1. Bayesian tests for small-scale experiments

The Bayesian approach to Boson Sampling validation ($\mathcal{V}_\mathrm{B}$), introduced in reference [31] and recently investigated also in reference [58], aims to identify the most likely between two alternative hypotheses, which model the multi-photon states under consideration. In particular, $\mathcal{V}_\mathrm{B}$ tests the Boson Sampling hypothesis (H_Q), which assumes fully indistinguishable n-photon states, against an alternative hypothesis (H_A) for the source that produces the measurement outcomes $\{x\}$. Equal probabilities are assigned to the two hypotheses prior to the experiment. Let us denote with $p_\mathrm{Q}(x_k)$ $\left(p_\mathrm{A}(x_k)\right)$ the scattering probability associated with the output state x_k for H_Q (H_A). The intuition is that, if H_Q is most suitable to model the experiment, it is more likely to collect events for which $p_\mathrm{Q}(x_k) > p_\mathrm{A}(x_k)$. The idea is made quantitative considering the confidence $P(\{x\}|H_\mathrm{hypo}) = \prod_{k=1}^{\mathcal{S}} p_\mathrm{hypo}(x_k)$ we assign to each hypothesis, with hypo being

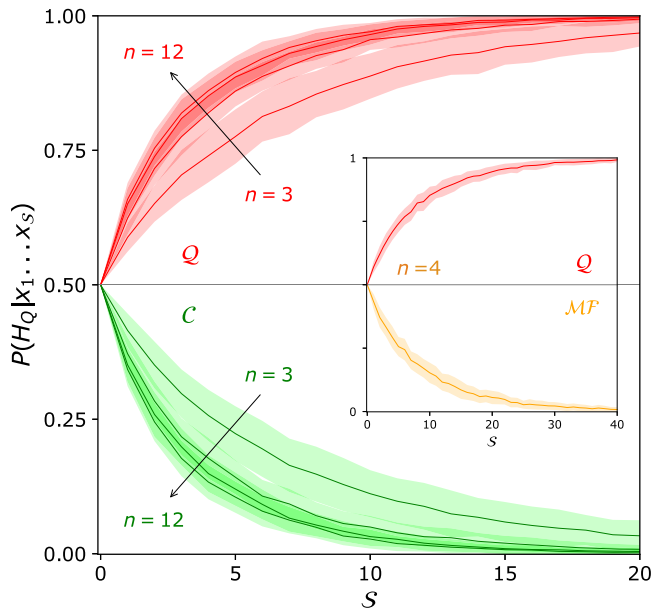

Figure 2. Confidence $P(H_Q|\{x\})$ of the Bayesian test to accept, as a correct Boson Sampling experiment, events that are sampled using distinguishable ($\mathcal{C}$, green) [20] and indistinguishable ($\mathcal{Q}$, red) [48] n-photon states from m-mode interferometers. Note how curves become steeper for increasing n ($n = 3, 6, 9, 12$ and $m = n^2$), making the test progressively more sample-efficient. Inset: Bayesian protocol applied to test $\mathcal{Q}$ against the mean-field sampler ($\mathcal{MF}$, orange) [25] for $n = 4$ photons and $m = n^2$. Curves are obtained by numerically sampling $10^4 n$-photon events, averaging over 50 random reshuffling of these events and over 100 different Haar-random unitary transformations (shaded regions: one standard deviation).

either Q or A. By applying Bayes' theorem, after $\mathcal{S}$ events we have

$$\frac{P(H_Q|\{x\})}{P(H_A|\{x\})} = \prod_{k=1}^{\mathcal{S}} \frac{p_Q(x_k)}{p_A(x_k)} \tag{1}$$

By combining equation (1) and $P(H_Q|\{x\}) + P(H_A|\{x\}) = 1$, it follows that our confidence in the hypothesis H_Q becomes $P(H_Q|\{x\}) = \frac{\chi_{\mathcal{S}}}{1+\chi_{\mathcal{S}}}$, with $\chi_{\mathcal{S}} = \prod_{k=1}^{\mathcal{S}} \frac{p_Q(x_k)}{p_A(x_k)}$.

This test requires the evaluation of permanents of $n \times n$ scattering matrices [50, 53], since

$$p_Q(x_k) = \left(\prod_{i=1}^{m} s_i! \prod_{i=1}^{m} t_i! \right)^{-1} \quad \mathrm{Per}^2(U_{S,T}) \tag{2}$$

where $U_{S,T}$ is the matrix obtained by repeating s_k (t_k) times the kth column (row) of $\mathcal{U}$, s_k (t_k) being the occupation number in the input (output) mode k ($\sum_{k=1}^{m} s_k = \sum_{k=1}^{m} t_k = n$). The presence of the permanent in equation (2) sets an upper limit to the number of photons that can be studied in practical applications [40–47]. Indeed, it is foreseeable that real-time monitoring or feedback-loop stabilization of quantum optics experiments will only have access to portable platforms with limited computational power. However, an interesting advantage of this validation protocol is its broad versatility, due to the absence of assumptions on the alternative distributions. Importantly, when applied to validate Boson Sampling with distinguishable photons, it requires very few measurements ($\mathcal{S} \sim 20$) for a reliable assessment. In figure 2, for instance, we numerically investigate its application as a function of sample size, extending previous simulations from $n = 3$ [31] to $n = (3, 6, 9, 12)$ and $m = n^2$.

Data for distinguishable (H_C) and indistinguishable (H_Q) photons were generated using exact algorithms, respectively by Aaronson and Arkhipov [20] and by Clifford and Clifford [48]. The analysis shows how the validation protocol becomes even more effective for increasing n, being it able to output a reliable verdict after only ~ 20 events. However, as mentioned, its power comes at the cost of being computationally inefficient in n. Also, it is not possible to preprocess $\mathcal{V}_B$ and store information for successive re-use, since its confidence depends on the specific $\mathcal{U}$ and sampled events, according to $p_Q(x_k)$. Hence, in the regime $n \sim 25$–35 [46, 47] it becomes rapidly harder to perform a validation in real time. Eventually, since classical supercomputers cannot assist quantum experiments in everyday applications, $\mathcal{V}_B$ becomes prohibitive from $n \sim 35$.

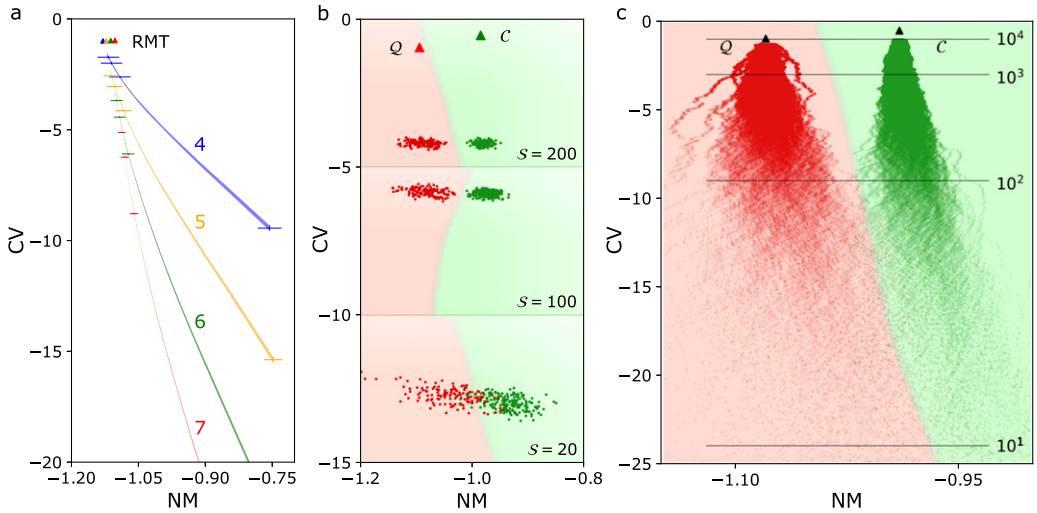

Figure 3. (a) Numerically simulated evolution of C-datasets in the NM-CV plane for an increasing sample size $\mathcal{S}$. Boson Sampling with $n = (4, 5, 6, 7)$ indistinguishable photons in $m = n^2$ modes, for $\mathcal{S} = 2, 3, 4, \ldots, 100$. For large $\mathcal{S}$, curves converge to the points (pyramids) predicted by random matrix theory (RMT) [30]. Points are averaged over 100 Haar-random unitaries, while error bars are displayed every 20 additional events. (b) Validation via statistical benchmark of numerically generated Boson Sampling data with indistinguishable photons (red) against ones with distinguishable photons ($\mathcal{C}$, green), for 100 different unitary transformations, $n = 8$ photons, $m = 64$ modes and $\mathcal{S} = 20, 100, 200$ events. Contour plots describe the confidence of (three different instances of) neural network binary classifiers, trained at different $\mathcal{S}$, from green (label: $\mathcal{C}$) to red (label: $\mathcal{Q}$). red and green pyramids identify the random matrix prediction from reference [30], for $S \to \infty$. Note how the clouds of points shrink for increasing $\mathcal{S}$, making the classification of experimental points progressively more reliable. (c) Same analysis as in (b), for 200 different unitary transformations, $n = 10$ photons, $m = 100$ modes and $\mathcal{S} = 1, 2, 3, \ldots 10^4$ events. The classifier is now trained on data from all $\mathcal{S}$. Numbers on the right indicate, approximately, the sample size at the corresponding height. Once trained, this classifier can be deployed to validate any other Haar-random experiment with the same (n, m) and any $\mathcal{S}$.

3.2. Statistical benchmark for large-scale experiments

In the previous section we described how the Bayesian test is effective in validating small- and mid-scale experiments with very few measurement events. However, the evaluation of permanents hinders its application for large n, be it due to too large scattering matrices or to the need for speed in real-time evaluations. To overcome this limitation, further validation protocols have been proposed in the last few years, to find a convenient compromise between predictive power and physical resources. All these approaches have their own strengths and limitations, and tackle the problem from different angles [16], e.g. using suppression laws [24–28], machine learning [29, 33] or statistical properties related to multi-particle interference [30]. In this section we will focus on the latter protocol, which arguably represents the most promising solution for the reasons we outline below.

Statistical benchmark with finite sample size. Validation based on the statistical benchmark ($\mathcal{V}_S$) looks at statistical features of the C-*dataset*, the set of two-mode correlators

$$C_{ij} = \langle \hat{n}_i \hat{n}_j \rangle - \langle \hat{n}_i \rangle \langle \hat{n}_j \rangle \tag{3}$$

where (i, j) are distinct output ports and $\hat{n}_i$ is the bosonic number operator. Two statistical features that are effective to discriminate states with indistinguishable and distinguishable photons are its normalized mean NM (the mean divided by n/m^2) and its coefficient of variation CV (the standard deviation divided by the mean). For any unitary transformation and input state we can retrieve a point in the plane (NM, CV), where alternative models tend to cluster in separate clouds located via random matrix theory (figure 3(a)) [30]. Validation based on $\mathcal{V}_S$ would then consist in (i) collecting a suitable number $\mathcal{S}$ of events, (ii) evaluating the experimental point (NM, CV) associated to the C_{ij} and (iii) identifying the cluster that the point is assigned to. For $\mathcal{S}$ sufficiently large, the point will be attributable with large confidence to only one of the models, thus ruling out the others (figure 3(b)).

$\mathcal{V}_S$ represents the state of the art for validation protocols that do not require the evaluation of permanents. Indeed, this approach has several advantages [39]: (a) it is computationally efficient (one only needs to compute two-point correlators), (b) it can reveal deviations from the expected behaviour (manifest in the NM-CV plane), (c) it makes more reliable predictions for larger n (clouds become more separate), (d) it is sample-efficient (clouds separate relatively early, after few measurements events). However, despite points (c, d) above, in actual conditions the experimental point is not always easy to validate. In fact, as mentioned in point (b), hardware imperfections and partial distinguishability make the point move away from the average route shown in figure 3(a). These issues can be addressed and mitigated by numerically

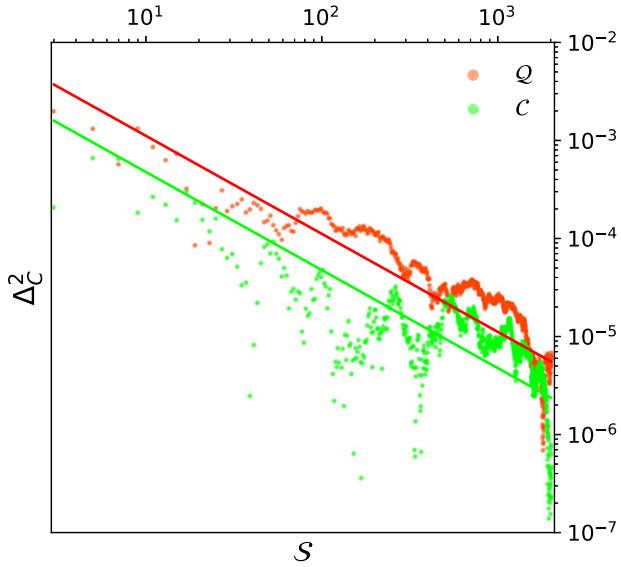

Figure 4. Log-log plot of the deviation $\Delta C^2_{\mathcal{S}} = |\mathbb{E}_U[\mathbb{E}_X(\widetilde{C}^2_{ij})] - \mathbb{E}_U[C^2_{ij}]| = |\mathbb{E}_U[\sigma^2_{ij}]/\mathcal{S}|$ from equation (4) as a function of the sample size $\mathcal{S}$. Data numerically generated to mimic experiments with $n = 4$ photons in $m = 16$ modes (green: distinguishable photons [20]; red: indistinguishable photons [48]). Averages are carried out over 500 Haar-random unitaries U and 500 different samples of size $\mathcal{S}$ (number of events) from each unitary, with fixed input state $(1, 1, 1, 1, 0, \ldots, 0)$. The linear fits to the different data sets exhibit the expected scaling $\propto \mathcal{S}^{-1}$.

generating, for a fixed sample size $\mathcal{S}$, clouds from unitary transformations that take these aspects into account. This intuition applies to all imperfections that can be described by an error model, for instance controlled by a set of parameters that quantify the noise level. Specifically, whenever it is possible to numerically generate events from a probability distribution that models these imperfections, we can use this data to train a classifier to recognize them in actual experiments. Relevant examples of error models include the above-mentioned partial distinguishability [16, 23–26, 60–62], unbalanced losses and fabrication imperfections in the optical elements of an interferometer [59].

As suggested in reference [39], and more closely investigated in figures 3(b) and (c), a convenient approach is to employ machine learning to assign experimental points to one of the two clouds, with a certain confidence level. Specifically, one can train a classifier with numerically generated data [20, 48] for a certain $(n, m, \mathcal{S})$, that can even include error models, and then deploy it for all applications in that regime. In this sense, $\mathcal{S}$ can be seen as the label of the model that can classify (validate) data for a given (n, m). This intuition can be extended to a classifier that is trained on data from multiple $\mathcal{S}$ [see figure 3(c)], which is likely more practical. For a fixed $\mathcal{S}$, the computational resources to sample events from a distribution given by n distinguishable (indistinguishable) photons scale polynomially [20] (exponentially [48]) in n. However, once trained, this classifier can be considered as an off-the-shelf tool that is readily applicable to validate multi-photon interference with no additional computational overhead, which is ideal for large-size experiments. In appendix B, we also discuss how such a classifier can even be combined with other protocols, which search the data for different distinctive structures, to boost its accuracy.

Finite-size effects in validation protocols. So far, we qualitatively discussed the role of a limited sample size for the validation of multi-photon quantum interference. To provide a more quantitative analysis of finite-size effects for the task of validation, and in particular for $\mathcal{V}_S$, in the following we study the scaling of the parameters involved in the above validation protocol with $\mathcal{S}$. The goal of this section is to elaborate on a standard test which should be implemented in all validation protocols, to guarantee their experimental feasibility.

Let us start by considering a fixed unitary circuit U, for which we calculate the correlators C_{ij} from equation (3). Such evaluation in principle assumes the possibility to collect an arbitrary number of measurement events. In practical applications, however, sample sizes will always be limited. Hence, finite-size effects play a role in the estimation of the above correlators. According to the central limit theorem, the correlator retrieved from the experimental data can be represented as $\widetilde{C}_{ij} = C_{ij} + X_{ij}$, where X_{ij} is a random number normally distributed with zero mean and variance $\sigma^2_{ij}\,\mathcal{S}^{-1}$. The σ^2_{ij} depend on the unitary evolution U and should either be evaluated from the data or be estimated using random matrix theory. Now, to infer, from noisy C-datasets [30], the centre of the cloud of points in the NM-CV plane, we need to average not only over the Haar measure, but also over X_{ij}.

Consequently, we have to assess the impact of finite-size effects on the estimate of the moments (NM, CV). First, since the noise induced by the finite sample size averages out, namely $\mathbb{E}_X(\widetilde{C}_{ij}) = C_{ij}$, we have that $\widetilde{\mathrm{NM}} = \mathrm{NM}$. The estimation of CV is a bit more subtle because we need to evaluate the mean of $\widetilde{C}_{ij}^{\,2}$. Since $\mathbb{E}_X(\widetilde{C}_{ij}^{\,2}) = C_{ij}^2 + \sigma_{ij}^2\,\mathcal{S}^{-1}$, then

$$\mathbb{E}_U[\mathbb{E}_X(\widetilde{C}_{ij}^2)] = \mathbb{E}_U[C_{ij}^2] + \frac{\mathbb{E}_U[\sigma_{ij}^2]}{\mathcal{S}} \tag{4}$$

and, hence, $|\widetilde{\mathrm{CV}}| > |\mathrm{CV}|$. Note that $\mathbb{E}_U[\mathbb{E}_X(\widetilde{C}_{ij}^2)]$ and $\mathbb{E}_X[\mathbb{E}_U(\widetilde{C}_{ij}^{\,2})]$ cannot be easily compared, since the latter involves averaging the distribution of X_{ij} over the unitary group. However, using the properties of the normal distribution under convex combinations, we can deduce that both orders of averaging yield approximately the same result (and the same scaling in $\mathcal{S}$), in particular once $\mathcal{S}$ is large and the distribution is concentrated close to its mean. Numerical simulations for $3 \leqslant n \leqslant 15$ and $m = n^2$ indeed confirm its validity (figure 4). Specifically, we observe that, upon averaging over different Haar-random unitaries with $\mathcal{S}$ events per realization, the deviation of the experimentally-measured $\widetilde{C}_{ij}^{\,2}$ from the analytically predicted values decreases as fast as $1/\mathcal{S}$. Hence, their estimation from finite-size data sets shows no exponential overhead that would hinder a practical application of the validation protocol.

4. Discussion

Validation of multi-photon quantum interference is expected to play an increasing role as the dimensionality of photonic applications increases, both in the number of photons and modes. To this end, and as notably emphasized by the race towards quantum advantage via Boson Sampling, it is necessary to define a set of requirements for a validation protocol to be meaningful. Ultimately, these requirements should allow to establish strong experimental evidence of quantum advantage that is accepted by the community within a jointly agreed framework.

In the present work, we implement such a program and describe a set of critical points that experimenters will need to agree upon in order to validate the operation of a quantum device. With the goal of building a solid framework for validation, we then discuss a practical approach that applies the most suitable state-of-the-art protocols in realistic scenarios. We report numerical analyses on the application of two key validation protocols, the Bayesian hypothesis testing and the statistical benchmark, with finite-size data, providing compelling evidence in support of this approach.

A clear and illustrative example for the above considerations is provided in appendix A, where we numerically studied the competition between a recent classical simulation algorithm and the statistical benchmark, respectively to counterfeit and to validate Boson Sampling, while they process an increasing number of measured output events. The analysis quantifies the general intuition that there must be a trade-off between speed and quality in approximate simulations of Boson Sampling. We also provide a formal analysis on the performance of the validation protocol with finite-size samples, showing that the estimation of relevant quantities converges fast to the predicted values. We expect that similar features will be crucial for larger-scale demonstrations and, as such, a key prerequisite to be investigated in all validation protocols.

Finally, in appendix B we introduce a novel approach to validation that can bring together the strengths of multiple protocols. This approach uses a meta-algorithm (AdaBoost) to combine protocols based on machine learning into a single validator with boosted accuracy. This strategy becomes more advantageous for a larger number of such protocols with comparable performance, as well as with very noisy data.

Acknowledgments

This work was supported by ERC Advanced Grant QU-BOSS (QUantum advantage via non-linear BOSon Sampling; Grant Agreement No. 884676); by the QuantERA ERA-NET Cofund in Quantum Technologies Project HiPhoP (High dimensional quantum Photonic Platform, Project ID 731473) and by project PRIN 2017 'Taming complexity via QUantum Strategies a Hybrid Integrated Photonic approach' (QUSHIP) Id. 2017SRNBRK. AB acknowledges support by the Georg H Endress foundation. MW is funded through Research Fellowship WA 3969/2-1 of the German Research Foundation (DFG). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement No. 801110 and the Austrian Federal Ministry of Education, Science and Research (BMBWF). It reflects only the author's view and the Agency is not responsible for any use that may be made of the information it contains.

Appendix A. Classical simulation and the role of sample size

To shed some light on the critical aspects of validation, and as a benchmark of the state of the art in this context, we now provide a qualitative analysis inspired by the metropolized independent sampling ($\mathcal{M}$), a recent algorithm to classically simulate Boson Sampling [47]. The idea behind $\mathcal{M}$ is reminiscent of the mean-field sampler ($\mathcal{MF}$) [25], an adversarial classical algorithm that was capable to hack one of the first validation protocols [32] using limited classical resources. In the race towards quantum computational supremacy, the introduction of $\mathcal{MF}$ has prompted the development of more sophisticated techniques to tackle classical simulations. For instance, besides the Bayesian test (see inset in figure 2), also the statistical benchmark is highly effective to validate Boson Sampling against $\mathcal{MF}$ (see figure A1(a)). For our scope, the key difference between the two algorithms is that, while for $\mathcal{MF}$ the quality of the simulation does not really change over time, $\mathcal{M}$ samples from a distribution that gets closer to $\mathcal{Q}$ the more events are evaluated (i.e. for a larger $\mathcal{S}$).

The goal of $\mathcal{M}$ is to generate a sequence of n-photon events $\{\mathrm{e}_i\}$ from a Markov chain that mimics the statistics of an ideal Boson Sampling experiment. Given a sampled event e_i, a new candidate event e_{i+1} is efficiently picked according to the probability distribution of distinguishable photons p_D, and accepted with probability

$$P(\mathrm{e}_i \to \mathrm{e}_{i+1}) = \min\left(1,\ \frac{p_\mathrm{I}(\mathrm{e}_{i+1})\, p_\mathrm{D}(\mathrm{e}_i)}{p_\mathrm{D}(\mathrm{e}_{i+1})\, p_\mathrm{I}(\mathrm{e}_i)}\right) \tag{5}$$

where $p_\mathrm{I}(\mathrm{e}_i)$ is the output probability corresponding to event e_i for indistinguishable photons. While the approach remains computationally hard, since it requires the evaluation of permanents [53, 63], the advantage is that only a limited number of them needs to be evaluated to output a new event, rather than the full distribution as in a brute-force approach. Ultimately, after a certain number of steps in the chain, $\mathcal{M}$ is guaranteed to sample close to the ideal Boson Sampling distribution p_I [64]. Hence, not only does the sample size $\mathcal{S}$ play a key role to improve the reliability of validation protocols, as shown in section 3, but it can be crucial also to increase the quality of the outcome of a classical simulation. This is a relevant point to keep in mind, even though $\mathcal{M}$ has since been surpassed by an algorithm that is both provably faster and exact [48, 49]. In fact, in future, novel classical algorithms might be developed [54] that depend on $\mathcal{S}$ more efficiently.

The aim of our present analysis is to investigate the role of the sample size in a validation of the samples generated by $\mathcal{M}$, via $\mathcal{V}_\mathrm{S}$. Indeed, a crucial issue in a hypothetical competition between $\mathcal{M}$ and $\mathcal{V}_\mathrm{S}$ concerns the number of events $\mathcal{S}$ available to accept or reject a data set. While larger sets provide deeper information to $\mathcal{V}_\mathrm{S}$ to identify fingerprints of quantum interference, on the other hand $\mathcal{M}$ approaches the target distribution p_I as more steps are made along the chain. However, in order to output a large number of events in time $\mathcal{T}$, $\mathcal{M}$ requires physical and computational resources that set a limit to the tractable dimension of the problem. We are then interested in the intermediate regime, the one relevant for experiments, to determine whether convergence is reached fast enough to mislead $\mathcal{V}_\mathrm{S}$. In the specific case of $\mathcal{M}$, we then need to look at the scaling in n of its hyper-parameters: *burn-in* (the number B_n of events to be discarded at the beginning of the chain) and *thinning* (the number T_n of steps to skip to reduce correlations between successive events). Eventually, the time required to classically simulate Boson Sampling will scale as $\mathcal{T} = \tau_\mathrm{p}\,(B_n + \mathcal{S}\ T_n)$, where τ_p is the time to evaluate a single scattering amplitude according to equation (5). Considering the estimate provided by the supercomputer Tianhe-2 [50], and for fixed ($\mathcal{T}$, $\mathcal{S}$), we find the constraint $B_n = \alpha\, n^{-2}\, 2^{-n}\, \mathcal{T} - \mathcal{S}\ T_n$ where $\alpha \sim c^{0.8782} \times 10^{11}$ and c is the number of processing nodes. If we assume $T_n = 100$ [47] for all n and $\mathcal{V}$, we get an estimate of the maximum B_n allowed by ($\mathcal{T}$, $\mathcal{S}$). The key issue is that this estimate does not guarantee that $\mathcal{M}$ achieves the target distribution fast enough, since B_n decreases (exponentially) in n. Moreover, the minimum B_n is expected to increase with n, since on average the Markov chain needs to explore more states before picking a good one.

To better clarify the above considerations, we simulate a competition between $\mathcal{M}$ and $\mathcal{V}_\mathrm{S}$ for $n = 10$ photons in $m = 100$ modes on figure A2. Data for distinguishable and indistinguishable photons were generated with exact algorithms, respectively by Aaronson and Arkhipov [20] and by Clifford and Clifford [48]. The analysis proceeds through five main steps: (1) randomly pick a unitary transformation $\mathcal{U}$ according to the Haar measure; (2) simulate the generation of $\mathcal{S}n$-particle output events; (3) extract the C-dataset from these $\mathcal{S}$ events; (4) evaluate the corresponding (NM, CV) point and plot it in figure A2(a); (5) repeat steps 1–4 200 times, to simulate as many different experiments. Upon completion, evaluate average and variance of $P_\mathcal{M}$ and plot them in figure A2(b). With this analysis, we get a quantitative intuition on how the confidence of a validation changes with $\mathcal{S}$, as does the quality of the classical simulation. Similar behaviour is found also for other choices of n and m. In particular, we observe how a stronger thinning (up to $T_{10} = 100$, as in reference [47]) is reflected in the quality of the simulation, where $\mathcal{M}$ behaves very

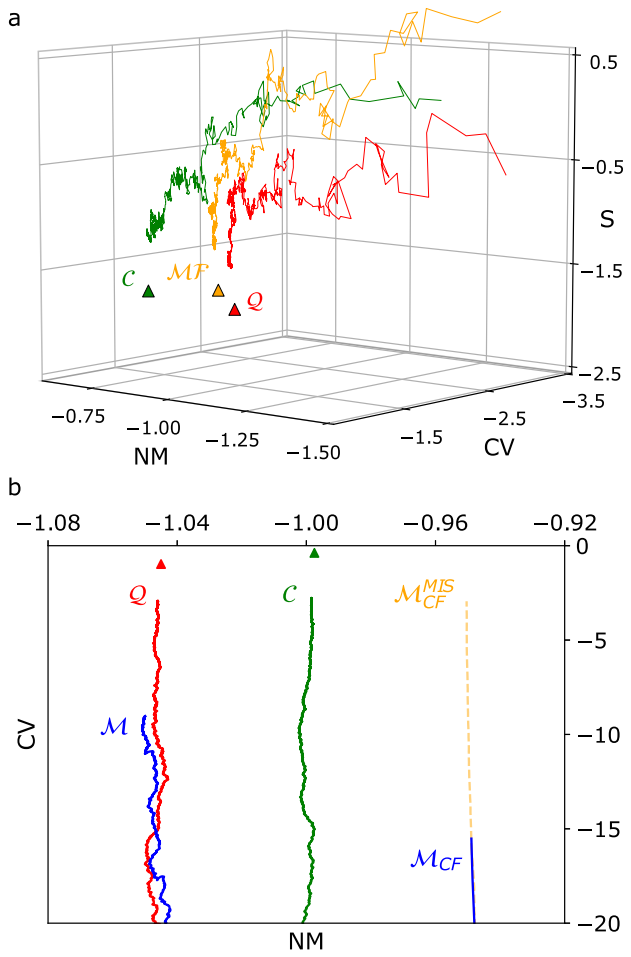

Figure A1. Numerically simulated evolution of C-datasets in the NM-CV-S space (a) and in the NM-CV plane (b) for an increasing sample size $\mathcal{S}$. (a) Boson Sampling with $n = 4$ indistinguishable photons ($\mathcal{Q}$, red) or distinguishable photons ($\mathcal{C}$, green) and mean-field sampler ($\mathcal{MF}$, orange) [25] in an $m = 16$-mode Haar-random transformation, for $\mathcal{S} = 10^4$ events. Pyramids identify the random matrix prediction for $S \to \infty$ [30]. (b) Boson Sampling with indistinguishable or distinguishable photons and metropolized independent sampling (MIS), with collision events ($\mathcal{M}$) or without ($\mathcal{M}_{\mathrm{CF}}^{\mathrm{MIS}}$, extracted from data in reference [47]; $\mathcal{M}_{\mathrm{CF}}$, data subset of $\mathcal{M}$) for $n = 20$ photons, $m = 400$ modes and up to $\mathcal{S} = 2 \times 10^4$ events. Curves without collision events (which can be resolved under stronger zoom) have a smoother evolution due to reduced fluctuations in the C-dataset. Note that the statistical benchmark captures the presence of collision events (in $\mathcal{Q}, \mathcal{C}, \mathcal{M}$), which have an impact on the statistics since the protocol probes two-particle processes.

similar to the ideal Boson sampler for small as well as for large sample sizes. Conversely, a faster $\mathcal{M}$ that trades quality for speed by computing fewer permanents ($T_{10} = 10, 30$) is more easily detectable by $\mathcal{V}_{\mathrm{S}}$. Constraints due to a speed vs quality compromise (figures 3(b)–(d)) define a generic scenario for a classical simulation which is run with a specific choice of $\mathcal{T}$ and $\mathcal{S}$.

Appendix B. Combining and boosting validation protocols

So far, all validation protocols have always been applied separately and independently. Certainly, this fact shows the multifaceted nature of this line of research, where effective solutions have been developed using very different strategies. Yet, it also reflects its somewhat fragmented condition, since each protocol does not benefit from potential insights provided by the others. This limitation becomes relevant in realistic scenarios with noise and finite data sets, since each validation protocol suits some task better than the others, with different degrees of sample efficiency and resilience.

In this section, we present a novel, synergistic approach to validation, which aims at combining the strengths of these protocols to form a joint, enhanced validator. Specifically, we focus on validation protocols that make use of machine learning, and propose to combine them with a meta-algorithm (AdaBoost [65]) that attempts an adaptive boosting of their individual performance. The output of AdaBoost is a weighted sum of the predictions of these learning algorithms ('weak learners'), which are asked, sequentially, to pay more attention to the instances that were incorrectly classified by the previous

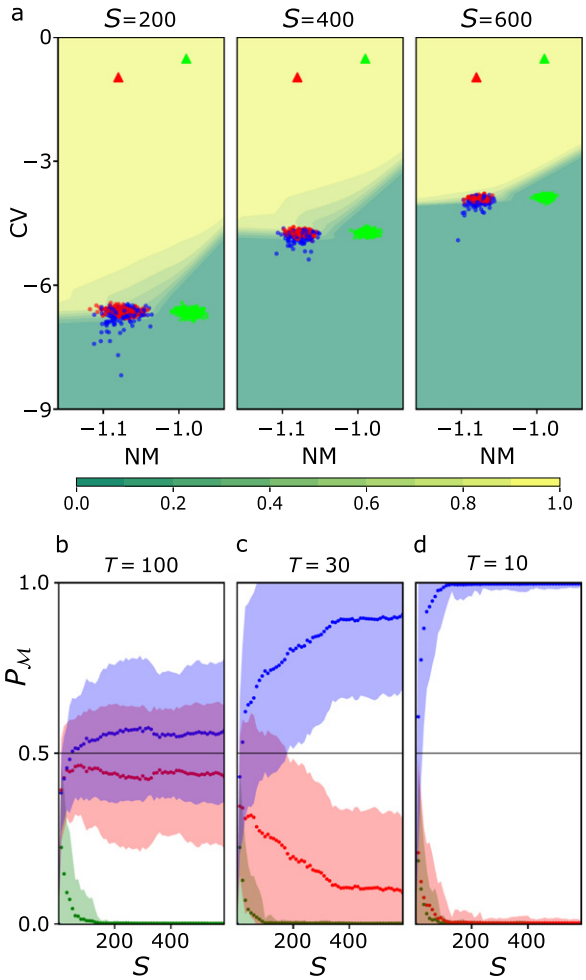

Figure A2. (a) Validation via statistical benchmark [30] of Boson Sampling with indistinguishable photons ($\mathcal{Q}$, red) against one with distinguishable photons ($\mathcal{C}$, green) and metropolized independent sampling [47] ($\mathcal{M}$, blue) with thinning $T = 100$ and burn-in $B = 0$, for 200 simulated experiments with different Haar-random unitary transformations, $n = 10$ photons, $m = 100$ modes and $\mathcal{S} = 200, 400, 600$ events. Contour plots describe the confidence of a neural network classifier, from green (low) to yellow (high), in labeling a point as $\mathcal{Q}$. Red and green pyramids identify the random matrix prediction from reference [30], for $S \to \infty$. (b)–(d) Confidence $P_{\mathcal{M}}$ of the same classifier in labeling (NM, CV) points generated by $\mathcal{M}$ as $\mathcal{M}$ (blue), $\mathcal{Q}$ (red) or $\mathcal{C}$ (green), for simulated experiments with $T = 100$ (b) from (a), $T = 30$ (c) and $T = 10$ (d). Values are averaged over all (NM, CV) points generated by $\mathcal{M}$, while shaded regions correspond to one standard deviation. Notice that in (b), with strong thinning, there still is a difference between $\mathcal{Q}$ and $\mathcal{M}$ data, though not significant due to larger fluctuations. Plots highlight the speed vs quality trade-off in classical simulations of Boson Sampling. See the main text for a step-by-step description of this analysis.

learners. As long as the performance of each learner is slightly better than chance, the classifier resulting from AdaBoost provably converges to a better validation protocol.

We numerically test this approach by combining two validation protocols that employ machine learning: the statistical benchmark $\mathcal{V}_{\mathrm{S}}$ [30] [equipped with a simple neural network classifier trained on numerically generated data, as in figures 3(b) and (c)] and the *visual assessment* $\mathcal{V}_{\mathrm{V}}$ [29], which uses dimensionality reduction algorithms and convolutional neural networks. Here we do not consider the Bayesian approach, since, in its current formulation, it does not fit the framework of machine learning. A schematic description of our proof-of-concept analysis, which we carry out for $n = 10$ and $m = 100$, is shown in figure A3.

Since $\mathcal{V}_{\mathrm{S}}$ requires fewer events than $\mathcal{V}_{\mathrm{V}}$ to validate ideal, noiseless experiments [20, 48], to perform this test we trained $\mathcal{V}_{\mathrm{S}}$ on data sets with a tunable amount noise, purposely assembled to be hard to validate. To this end, samples ($\mathcal{S} = 2 \times 10^3$) for 500 Haar-random unitary transformations were constructed by sampling with a certain probability p (or $1 - p$) from a Boson sampler with fully indistinguishable (or distinguishable) photons. This probability p was then varied in time, to simulate, for instance, a periodic drift in the synchronization of the input photons. As expected with these settings, we find that AdaBoost maintains the original accuracy of $\mathcal{V}_{\mathrm{S}}$ and $\mathcal{V}_{\mathrm{V}}$ when applied to, respectively, batches of $\mathcal{V}_{\mathrm{S}}$ and $\mathcal{V}_{\mathrm{V}}$ that are already highly accurate. This is mainly due to complexity of these classifiers, which are already strong learners and, hence, hard to enhance by AdaBoost. Analogous results are found with mixed batches of $\mathcal{V}_{\mathrm{S}}$ and $\mathcal{V}_{\mathrm{V}}$, for which AdaBoost returns a joint classifier that practically focuses on the most accurate one in the

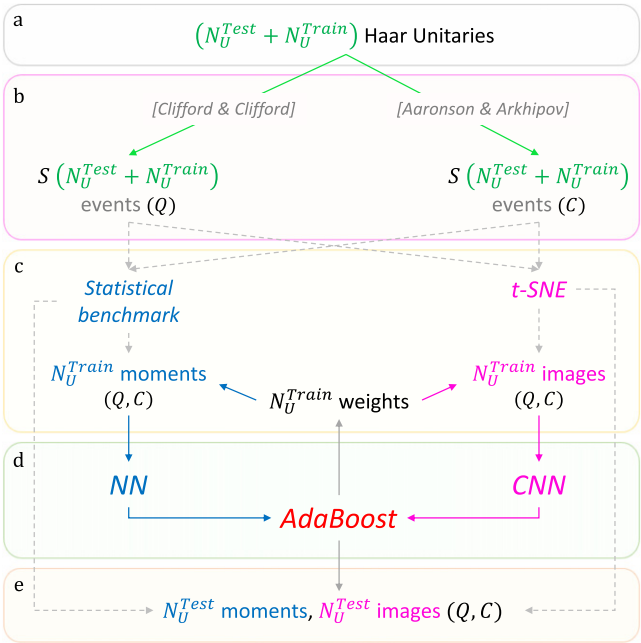

Figure A3. Machine learning techniques, such as AdaBoost [65], can combine individual validation protocols to boost the overall accuracy. A schematic overview of the approach is shown in this figure. Experiments associated with $N_U^{\mathrm{Test}} + N_U^{\mathrm{Train}}$ Haar-random unitary transformations (a) are simulated using exact algorithms to numerically sample $\mathcal{S}$ events from quantum [48] (Q) and classical [20] (C) Boson Sampling (b). (c) N_U^{Train} sets of $\mathcal{S}$ events for both Q and C are pre-processed by a collection of validation protocols. Here, we considered the statistical benchmark $\mathcal{V}_{\mathrm{S}}$ [30] and the visual assessment $\mathcal{V}_{\mathrm{C}}$ [29], which produced N_U^{Train} input data in the form of, respectively, pairs of moments (NM, CV) and images (using t-SNE [66] for dimensionality reduction). (d) Each protocol has its own classifier, in this case a neural network (NN) for $\mathcal{V}_{\mathrm{S}}$ and a convolutional neural network (CNN) for $\mathcal{V}_{\mathrm{V}}$, respectively. These classifiers are then applied sequentially on the same input data, iteratively adjusting their weights (AdaBoost) to focus on misclassifed data. (e) The resulting, joint protocol has higher accuracy on test data than each individual classifier.

set. A different result is obtained, instead, by combining several weak $\mathcal{V}_{\mathrm{V}}$, for which we purposely spoil the training of the convolutional neural network (accuracy $A \sim 51\%$ instead of $A \sim 98\%$) by reducing the number of training epochs. In this case, AdaBoost does in fact enhance the accuracy of $\mathcal{V}_{\mathrm{V}}$ up to $A \sim 57\%$.

In future, we expect that this approach will prove useful in non-ideal conditions with experimental noise, where validation protocols do not operate in the ideal settings where they were conceived. Furthermore, the above analyses can show larger boosts if applied to actual experiments that involve structured (non-Haar-random) interferometers, for which protocols such as $\mathcal{V}_{\mathrm{S}}$ and $\mathcal{V}_{\mathrm{V}}$ can have lower accuracies and different behaviors. Finally, still in non-ideal settings, more favorable boosts can be obtained if new validation protocols are developed that are as sample-efficient as $\mathcal{V}_{\mathrm{S}}$.

ORCID iDs

Fulvio Flamini https://orcid.org/0000-0003-4999-2840
Nicolò Spagnolo https://orcid.org/0000-0003-3471-2252
Fabio Sciarrino https://orcid.org/0000-0003-1715-245X

References

[1] Harrow A W and Montanaro A 2017 *Nature* **549** 203–9
[2] Arute F *et al* 2019 *Nature* **574** 505–10
[3] Aaronson S and Arkhipov A 2011 The computational complexity of linear optics *Proc. of the 43rd Annual ACM Symp. on Theory of Computing* (San Jose, California) (USA New York: Association for Computing Machinery) pp 333–42
[4] Terhal B M and DiVincenzo D P 2004 Adaptive quantum computation, constant depth quantum circuits, and Arthur Merlin games *Quantum Inf. Comput.* **4** 134–45
[5] Bremner M J, Jozsa R and Shepherd D J 2010 *Proc. R. Soc.* A **467** 459–72
[6] Boixo S, Isakov S V, Smelyanskiy V N, Babbush R, Ding N, Jiang Z, Bremner M J, Martinis J M and Neven H 2018 *Nat. Phys.* **14** 595–600
[7] Inagaki T *et al* 2016 *Science* **354** 603–6
[8] McMahon P L *et al* 2016 *Science* **354** 614–7
[9] Goto H 2016 *Sci. Rep.* **6** 21686
[10] Lechner W, Hauke P and Zoller P 2015 *Sci. Adv.* **1** e1500838

[11] Johnson M W *et al* 2011 *Nature* **473** 194–8
[12] Puri S, Andersen C K, Grimsmo A L and Blais A 2017 *Nat. Commun.* **8** 15785
[13] Fitzsimons J F and Kashefi E 2017 *Phys. Rev.* A **96** 012303
[14] Eisert J, Hangleiter D, Walk N, Roth I, Markham D, Parekh R, Chabaud U and Kashefi E 2019 **2** 382–90
[15] Flamini F, Spagnolo N and Sciarrino F 2018 *Rep. Prog. Phys.* **82** 016001
[16] Walschaers M 2020 *J. Phys.* B **53** 4
[17] Shin S W, Smith G, Smolin J A and Vazirani U 2014 How 'quantum' is the d-wave machine? (arXiv:1401.7087)
[18] Brod D J, Galvão E F, Crespi A, Osellame R, Spagnolo N and Sciarrino F 2019 *Adv. Photonics* **1** 034001
[19] Mayer K, Tichy M C, Mintert F, Konrad T and Buchleitner A 2011 *Phys. Rev.* A **83** 062307
[20] Aaronson S and Arkhipov A 2014 Bosonsampling is far from uniform *Quantum Inf. Comput.* **14** 1383–423
[21] Liu K, Lund A P, Gu Y J and Ralph T C 2016 *J. Opt. Soc. Am.* B **33** 1835–41
[22] Wang S T and Duan L M 2016 Certification of Boson Sampling devices with coarse-grained measurements (arXiv:1601.02627)
[23] Shchesnovich V S 2016 *Phys. Rev. Lett.* **116** 123601
[24] Tichy M C, Tiersch M, De Melo F, Mintert F and Buchleitner A 2010 *Phys. Rev. Lett.* **104** 220405
[25] Tichy M C, Mayer K, Buchleitner A and Molmer K 2014 *Phys. Rev. Lett.* **113** 020502
[26] Dittel C, Dufour G, Walschaers M, Weihs G, Buchleitner A and Keil R 2018 *Phys. Rev. Lett.* **120** 240404
[27] Dittel C, Keil R and Weihs G 2017 *Quantum Sci. Technol.* **2** 1
[28] Crespi A 2015 *Phys. Rev.* **91** 013811
[29] Flamini F, Spagnolo N and Sciarrino F 2019 *Quantum Sci. Technol.* **4** 024008
[30] Walschaers M, Kuipers J, Urbina J D, Mayer K, Tichy M C, Richter K and Buchleitner A 2016 *New J. Phys.* **18** 032001
[31] Bentivegna M *et al* 2014 *Int. J. Quantum Inf.* **12** 1560028
[32] Carolan J *et al* 2014 *Nat. Photonics* **8** 621–6
[33] Agresti I, Viggianiello N, Flamini F, Spagnolo N, Crespi A, Osellame R, Wiebe N and Sciarrino F 2019 *Phys. Rev.* X **9** 011013
[34] Spagnolo N *et al* 2014 *Nat. Photonics* **8** 615–20
[35] Viggianiello N *et al* 2018 *New J. Phys.* **20** 033017
[36] Carolan J *et al* 2015 *Science* **349** 711
[37] Crespi A *et al* 2016 *Nat. Commun.* **7** 10469
[38] Wang H *et al* 2018 *Phys. Rev. Lett.* **120** 230502
[39] Giordani T *et al* 2018 *Nat. Photonics* **12** 173–8
[40] He Y *et al* 2017 *Phys. Rev. Lett.* **118** 190501
[41] Wang H *et al* 2017 *Nat. Photonics* **11** 361–5
[42] Wang X L *et al* 2016 *Phys. Rev. Lett.* **117** 21
[43] Wang L J *et al* 2017 *Phys. Rev.* A **95** 012301
[44] Viggianiello N, Flamini F, Bentivegna M, Spagnolo N, Crespi A, Brod D J, Galvão E F, Osellame R and Sciarrino F 2018 *Sci. Bull.* **63** 22
[45] Paesani S *et al* 2019 *Nat. Phys.* **15** 925–9
[46] Wang H *et al* 2019 *Phys. Rev. Lett.* **123** 250503
[47] Neville A, Sparrow C, Clifford R, Johnston E, Birchall P M, Montanaro A and Laing A 2017 *Nat. Phys.* **13** 1153–7
[48] Clifford P and Clifford R 2018 The classical complexity of Boson sampling *Proc. of the 29th Annual ACM-SIAM Symp. on Discrete Algorithms SODA'18* ed A Czumaj (Philadelphia, PA: Society for Industrial and Applied Mathematics) pp 146–55
[49] Clifford P and Clifford R 2005 Faster classical Boson Sampling (arXiv:quant-ph/2005.04214v1)
[50] Wu J, Liu Y, Zhang B, Jin X, Wang Y, Wang H and Yang X 2018 *Natl Sci. Rev.* **5** 715–20
[51] Hangleiter D, Kliesch M, Eisert J and Gogolin C 2019 *Phys. Rev. Lett.* **122** 210502
[52] Markov I L, Fatima A, Isakov S and Boixo S 2018 Quantum supremacy is both closer and farther than it appears (arXiv:1807.10749)
[53] Valiant L 1979 *Theor. Comput. Sci.* **8** 189–201
[54] Liu Y *et al* 2019 Sample caching Markov chain Monte Carlo approach to Boson Sampling simulation (arXiv:1907.08077)
[55] Lowd D and Meek C 2005 *Adversarial Learning KDD* (New York: Association for Computing Machinery) pp 641–47
[56] Brugger J 2018 Quantum annealing and disorder: a statistical analysis *Master's Thesis* Albert-Ludwigs-Universität Freiburg
[57] Given that the validators have a non-zero probability of false positives or negatives the criteria for validation need to be generalized. In this context a device is said to be validated if the probability that a valid device passes (or fails) the tests is greater by a fixed fraction than the probabilities that would be predicted by all available alternative hypotheses.
[58] Dai Z, Liu Y, Xu P, Xu W X, Yang X J and Wu J J 2020 A Bayesian validation approach to practical boson sampling *Sci. China Phys. Mech. Astron.* **63** 250311
[59] Flamini F, Spagnolo N, Viggianiello N, Crespi A, Osellame R and Sciarrino F 2017 Benchmarking integrated linear-optical architectures for quantum information processing *Sci. Rep.* **7** 15133
[60] Dufour G, Brünner T, Rodríguez A and Buchleitner A 2020 Many-body interference in bosonic dynamics (arXiv:2005.07239)
[61] Giordani T *et al* 2020 *New J. Phys.* **22** 043001
[62] Renema J J, Menssen A, Clements W R, Triginer G, Kolthammer W S and Walmsley I A 2018 *Phys. Rev. Lett.* **120** 220502
[63] Scheel S 2008 *Acta Phys. Slovaca* **58** 675
[64] Brooks S P and Roberts G O 1998 *Stat. Comput.* **8** 319–35
[65] Freund Y and Schapire R E 1997 *J. Comput. Syst. Sci.* **55** 119–39
[66] van der Maaten L J P and Hinton G E 2008 *J. Mach. Learn. Res.* **9** 2579–605