Pattern Recognition Techniques for Boson Sampling Validation

Iris Agresti,[1] Niko Viggianiello,[1] Fulvio Flamini,[1] Nicolò Spagnolo,[1] Andrea Crespi,[2,3] Roberto Osellame,[2,3] Nathan Wiebe,[4] and Fabio Sciarrino[1]

[1]*Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 5, I-00185 Roma, Italy*
[2]*Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche (IFN-CNR), Piazza Leonardo da Vinci, 32, I-20133 Milano, Italy*
[3]*Dipartimento di Fisica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, I-20133 Milano, Italy*
[4]*Station Q Quantum Architectures and Computation Group, Microsoft Research, Redmond, Washington 98052, USA*

(Received 24 January 2018; revised manuscript received 7 November 2018; published 23 January 2019)

The difficulty of validating large-scale quantum devices, such as boson samplers, poses a major challenge for any research program that aims to show quantum advantages over classical hardware. Towards this aim, we propose a novel data-driven approach, wherein models are trained to identify common pathologies using unsupervised machine-learning methods. We illustrate this idea by training a classifier that exploits K-means clustering to distinguish between boson samplers that use indistinguishable photons from those that do not. We tune the model on numerical simulations of small-scale boson samplers and then validate the pattern-recognition technique on larger numerical simulations as well as on photonic chips in both traditional boson-sampling and scatter-shot experiments. The effectiveness of such a method relies on particle-type-dependent internal correlations present in the output distributions. This approach performs substantially better on the test data than previous methods and underscores the ability to further generalize its operation beyond the scope of the examples that it was trained on.

DOI: 10.1103/PhysRevX.9.011013

Subject Areas: Computational Physics, Photonics, Quantum Information

I. INTRODUCTION

There has been a flurry of interest in quantum science and technology in recent years that has been focused on the transformative potential that quantum computers have for cryptographic tasks [1], machine learning [2,3], and quantum simulation [4,5]. While existing quantum computers fall short of challenging their classical brethren for these tasks, a different goal has emerged that existing quantum devices could address: namely, testing the extended Church-Turing thesis. The extended Church-Turing thesis is a widely held belief that asserts that every physically reasonable model of computing can be efficiently simulated using a probabilistic Turing machine. This statement is, of course, controversial, since, if it were true, then quantum computing would never be able to provide exponential advantages over classical computing. Consequently, providing evidence that the extended Church-Turing thesis is wrong is more philosophically important than the ultimate goal of building a quantum computer.

Various intermediate computing models have been proposed in the past few years that promise to be able to provide evidence of a quantum computational supremacy, namely, the regime where a quantum device starts outperforming its classical counterpart in a specific task. Such models mostly belong to the category of sampling problems, i.e., simulating the distribution sampled from a quantum system, that is believed to be classically hard to compute. These include quantum circuits with commuting gates [6–8], quantum simulators with fully certifiable final states [9], and quantum random circuits [10].

A significant step in this direction has been achieved, in particular, by Aaronson and Arkhipov [11] with the formal definition of a dedicated task known as boson sampling. This task is a computational problem that consists in sampling from the output distribution of n indistinguishable bosons evolved through a linear unitary transformation. This problem has been shown to be classically intractable (even approximately) under mild complexity-theoretic assumptions. Indeed, the existence of a classical efficient algorithm to perform boson sampling would imply the collapse of the polynomial hierarchy to the third level [11]. Such a collapse is viewed among many computer scientists as being akin to violating the laws of thermodynamics. Thus, demonstrating that a quantum device can efficiently perform boson sampling is powerful evidence against the

extended Church-Turing thesis. Furthermore, the simplicity of boson sampling has already allowed experiments at a small scale with different photonic platforms [12–25], and also alternative approaches have been proposed, e.g., exploiting trapped ions [26] or applying random gates in superconducting qubits [10].

Despite the fact that boson sampling is within our reach, a major caveat remains. The measurement statistics for boson samplers are intrinsically exponentially hard to predict. This difficulty implies that, even if someone manages to build a boson sampler that operates in a regime beyond the reach of classical computers, then the experimenter needs to provide evidence that their boson sampler functions properly for the argument against the extended Church-Turing thesis to be convincing. This task is not straightforward, in general, for large quantum systems [27–30], and it represents a critical point for all the above-mentioned platforms seeking a first demonstration of quantum supremacy. Indeed, to conclusively certificate a sampler using a conventional metric, such as a cross-entropy, it is necessary to estimate the probability that the true boson sampler yields the outcome probability. However, computing such a probability efficiently would imply BQP = #P, which is totally implausible from a complexity-theoretic standpoint even given a quantum computer. Hence, one could consider providing progressively more stringent tests able to exclude relevant alternative error models. A first approach to ensure quantum interference could involve testing pairwise mutual indistinguishability by two-photon Hong-Ou-Mandel experiments [31]; however, such a method fails to completely characterize multiphoton interference [32]. While techniques exist that use likelihood ratios to validate [17,33,34], they work only for small systems. Other existing techniques exploit statistical properties of bosonic states [18,35–39] or symmetries of certain boson samplers [21,40–45]; however, these methods are much more limited in scope.

In this article, we devise a prototypical methodology based on machine-learning techniques to detect known types of malfunctions occurring in a quantum hardware performing sampling tasks. We apply the test here in the context of boson sampling; however, the intuition can be reformulated for other problems. This method compares features of the collected data sample with those of a second one obtained from a reference distribution, evaluating their compatibility. By generating the reference sample from an efficiently computable distribution corresponding to a well-defined pathology of the problem, it is then possible to exclude that such pathology is observed in the measured sample. More specifically, building on results of Wang and Duan [46], we devise a compatibility test between a trusted boson sampler and an untrusted device that looks at data structure in a suitable space (see Fig. 1). We then test experimentally our method on both traditional boson sampling and scatter-shot boson sampling [19,20], showing

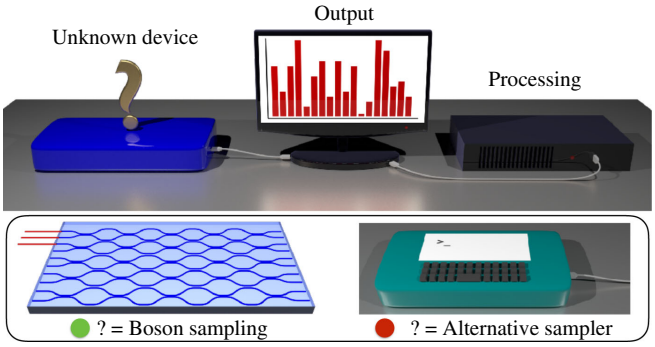

FIG. 1. Validation of boson-sampling experiments. An agent has to discriminate whether a finite sample obtained from an unknown device has been generated by a quantum device implementing the boson-sampling problem or by an alternative sampler.

that the algorithm is able to identify relevant pathologies in the measured samples. Finally, we provide a physical insight on the mechanism behind the functioning of the proposed clustering-based method in the investigated boson-sampling framework. Thanks to their versatility and their capability to operate without an in-depth knowledge of the physical system under investigation, clustering techniques may prove effective in a scope even broader than boson sampling [6–10].

II. BOSON SAMPLING AND ITS VALIDATION

Before going into detail about our approach, we need to discuss the boson-sampling problem at a more technical level. Boson sampling is a computational problem [11] that corresponds to sampling from the output probability distribution obtained after the evolution of n identical, i.e., indistinguishable, bosons through an m-mode linear transformation. Inputs of the problem are a given $m \times m$ Haar-random unitary matrix U, describing the action of the network on the bosonic operators according to the input-output relation $a_i^\dagger = \sum_j U_{i,j} b_j^\dagger$, and a mode occupation list $S = \{s_1, \ldots, s_m\}$, where s_i is the number of bosons on input mode i, being $\sum_i s_i = n$. For $m \gg n^2$ and considering the case where at most one photon is present in each input ($s_i = \{0, 1\}$) (collision-free scenario), sampling, even approximately, from the output distribution of this problem is classically hard. Indeed, in this regime for (n, m) the probability of a collision event becomes negligible, and, thus, the only relevant subspace is the collision-free one [11,16]. The complexity, in n, of the known classical approaches to perform boson sampling relies on the relationship between input-output transition amplitudes and, therefore, on the calculation of permanents of complex matrices, which is #P hard [47]. More specifically, given an input configuration S and an output configuration T, the transition amplitude $\mathcal{A}_U(S, T)$ between these two states is obtained as $\mathcal{A}_U(S, T) = \mathrm{per}(U_{S,T})/(s_1! \ldots s_n! t_1! \ldots t_n!)^{1/2}$,

TABLE I. Summary of the available protocols to validate quantum interference (Q) against experiments with distinguishable photons (D) or mean-field states (MF). Ideal validations should be reliable and efficient, meaning that they should not require resources exponential in the size of the problem, neither computational (e.g., the evaluation of permanents) nor physical (e.g., in the number of samples N). Also, they could provide insights on the multiphoton dynamics for a given transformation, as well as being applicable to conditions other than Q, D, or MF. LR, likelihood ratio [17,20,23–25]; ZTL, zero-transmission law (or suppression law) [21,34,40–45]; Bayesian [23–25,33,34]; bunching [18,35]; CG, coarse-graining [46]; $n = m$ [38]; statistical benchmark [36,39].

	Rules out:		*Does not require:*		*Gives insights on:*		*Proved in:*	
Test	D	MF	U	N large	dynamics	other $\hat{U}$	Theory	Exp
Likelihood ratio	✓	?		✓		✓	✓	✓
Bayesian test	✓	?		✓		✓	✓	✓
Bunching	✓		✓	?				✓
Suppression laws	✓	✓	✓				✓	✓
Coarse-grained	✓	✓	✓	✓	?	✓	✓	
$n = m$	✓	✓	✓				✓	
Statistical	✓	✓	✓	✓	✓	✓	✓	✓
This work	✓	✓	✓	✓	✓	✓	✓	✓

where $\mathrm{per}(U_{S,T})$ is the permanent of the $n \times n$ matrix $U_{S,T}$ obtained by selecting columns and rows of U according to the occupation lists S and T [48].

From a practical perspective, an essential aspect of any algorithm is that of verifying the correctness of its outputs. Verification can be trivial, as for factoring [1], or exponentially hard, as for boson sampling. In the latter case, which likely involves the evaluation of permanents, this stage is commonly referred to as *certification*. A similar goal, which aims to reduce the required physical resources in view of large-scale applications, is instead that of *validation*. In this case, one does not exactly attempt to verify the correctness of an outcome but, rather, to exclude known undesired models for the system that produced it. This approach has the advantage of being able to rule out many of the most likely pathologies that a boson sampler can face, without impacting the algorithm's practicality.

Thus, due to the complexity of evaluating the permanent, it is necessary to identify methods that do not require the calculation of input-output probabilities to validate the functioning of the device. Furthermore, in a quantum supremacy regime, the number of input-output combinations becomes very large, since it scales as $\binom{m}{n}$. Hence, it is also necessary to develop suitable techniques that are tailored to deal with a large amount of data. In Table I, we report a summary of the currently developed techniques, highlighting their main features and performances, in comparison with the method proposed in this article.

III. PATTERN-RECOGNITION TECHNIQUES FOR VALIDATION

In the regime where a boson-sampling device is expected to outperform its classical counterpart, the validation problem has inherently to deal with the exponential growth of the number of input-output combinations. A promising tool in this context is provided by the field of machine learning, which studies how a computer can acquire information from input data and learn to make data-driven predictions or decisions [49]. Significant progress has been achieved in this area over the past few years [50,51]. One of its main branches is represented by unsupervised machine learning, where dedicated algorithms have to find an inner structure in an unknown data set. One of the main unsupervised learning approaches is clustering, where data are grouped in different classes according to collective properties recognized by the algorithm [52]. Since this approach is designed to identify hidden patterns in a large amount of data, clustering techniques are promising candidates to be applied for the boson-sampling validation problem.

Let us discuss the general scheme of the proposed validation method based on pattern-recognition techniques. This approach allows us to employ various clustering methods within the protocol, which allow us to choose the method that optimizes the performance on the training data. Given two samples obtained respectively from a *bona fide* boson sampler, that is, a trusted device, and a boson sampler to be validated, the sequence of operations consists in (i) finding a cluster structure inside the data belonging to the first sample, (ii) once the structure is completed, organizing the data of the second sample by following the same structure of the previous set, and (iii) performing a χ^2 test on the number of events per cluster for the two independent samples. The χ^2 variable is evaluated as $\chi^2 = \sum_{i=1}^{N_c} \sum_{j=1}^{2} [(N_{ij} - E_{ij})^2 / E_{ij}]$, where index j refers to the samples and index i to the N_c clusters, N_{ij} is the number of events in the ith cluster belonging to the jth sample, and E_{ij} is the expected value of observed events belonging to the jth sample in the ith cluster $E_{ij} = N_i N_j / N_c$, with $N_i = \sum_{j=1}^{2} N_{ij}$, $N_j = \sum_{i=1}^{N_c} N_{ij}$, and $N = \sum_{i=1}^{N_c} \sum_{j=1}^{2} N_{ij}$. If the null hypothesis of the two samples being drawn from the same probability

distribution is correct, the evaluated variable must follow a χ^2 distribution with $\nu = N_c - 1$ degrees of freedom (d.o.f.). This scheme can be applied by adopting different metric spaces and different clustering techniques. Concerning the choice of the metric, both 1-norm and 2-norm distances can be employed as the distance d between two Fock states Ψ and Φ, namely, $d = L_1 = \sum_{i=1}^{M} |\Psi_i - \Phi_i|$ or $d = L_2 = \sqrt{\sum_{i=1}^{M} |\psi_i - \phi_i|^2}$, with Ψ_i and Φ_i being, respectively, the occupation numbers of Ψ and Φ in the ith mode.

IV. ADOPTED CLUSTERING TECHNIQUES

Several clustering methods are employed within our validation scheme: (a) a recent proposal by Wang and Duan [46], whose concept is shown in Fig. 2, and two unsupervised machine-learning techniques, (b) agglomerative hierarchical clustering and (c) K-means clustering. Two variations of the latter approach are also examined, to increase the strength of our model. A short description of each adopted method follows briefly.

(a) The protocol proposed by Wang and Duan [46], and hereafter named *bubble clustering*, determines the inner cluster structure of a sample by (i) sorting in decreasing order the output events according to their frequencies, (ii) choosing the observed state with the highest frequency as the center of the first cluster, (iii) assigning to such a cluster all the states belonging to the sample whose distance d from its center is smaller than a cutoff radius ρ_i, and (iv) iterating the procedure with the remaining states until all the observed events are assigned.

(b) Hierarchical clustering, in its bottom-up version, starts by assigning each observed event to a separate class. Then, the two nearest ones are merged to form a single cluster. This grouping step is iterated, progressively reducing the number of classes. The agglomeration stops when the system reaches a given halting condition predetermined by the user. In the present case, the algorithm halts when no more than 1% of the observed events are included in some cluster containing less than five events (see Supplemental Material [53]). All of these smallest clusters are considered as outliers and removed from the structure when performing the χ^2 test. The distance between two clusters is evaluated as the distance between their centroids. The centroid of a cluster is defined as the point that minimizes the mean distance from all the elements belonging to it.

(c) K means is a partitioning clustering algorithm where the user has to determine the number of classes (k) [54–56]. With this method, the starting points for centroid coordinates are chosen randomly. Then, two operations are iterated to obtain the final cluster structure that are selecting elements and moving centroids. The first one consists in assigning each observed event to the cluster whose centroid has the smallest distance from it. Then, once the k clusters are completed, the centroid of each cluster is moved from the previous position to an updated one, given by the mean of the element coordinates. These two operations are repeated until the structure is stable. Given a set of k centroids $(c_1, \ldots, c_k)$ made of $(N_1, \ldots, N_k)$ elements $(e_{11}, \ldots, e_{1n_1}, \ldots, e_{k1}, \ldots, e_{kn_k})$, where $\sum_{i=1}^{k} N_i = n$, the operations of selecting elements and moving the centroids minimize the objective

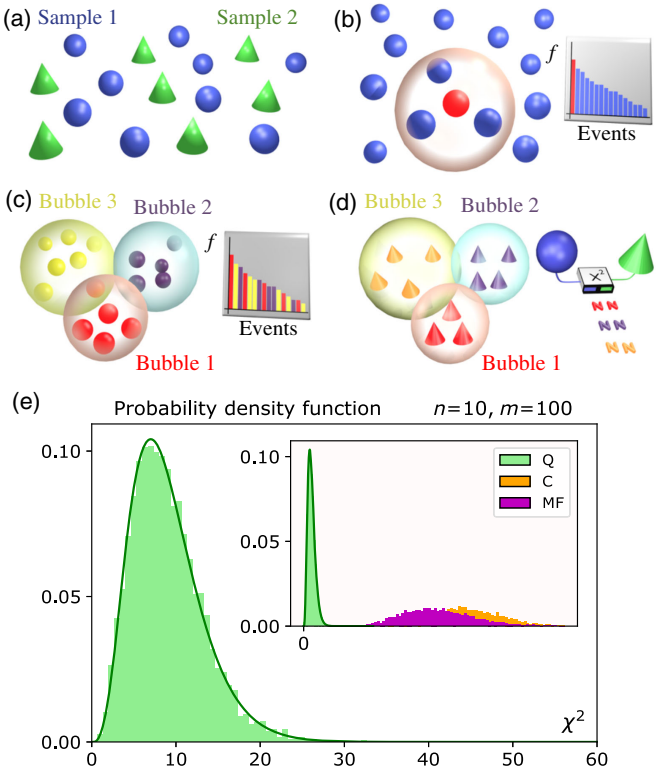

FIG. 2. Pattern-recognition techniques for validation. (a) A sample is drawn from each of the two boson samplers to be compared. The events belonging to one of the two samples are partitioned according to the criteria of the pattern-recognition technique. (b),(c) Bubble clustering sorts the events according to their observation frequency, and the state with the highest frequency is chosen as the center of the first cluster: All events with a distance from the center smaller than a cutoff radius ρ_1 are included in this cluster (b). Then, starting from the unassigned events, this procedure is iterated until all of the observed events are included in some bubble. At this point, each cluster is characterized by a center and a radius (c). (d) The observed events belonging to the second sample are classified by using the structure tailored from the first sample: Each event belongs to the cluster with the nearest center. (e) A χ^2 test with $\nu = N_{\text{bubbles}} - 1$ d.o.f. is performed (here using K means) to compare the number of events belonging to each of the two samples [green, quantum (Q) with indistinguishable photons; orange, classical (C) with distinguishable photons; purple, mean-field state] by using the obtained cluster structure. This variable quantifies the compatibility between the samples.

function $\frac{1}{N_c}\sum_{j=1}^{k}\sum_{i=1}^{N_k} d(e_{ij}, c_j)$. Several trials are made to determine the optimal number of clusters, showing that the performance of the test improves for higher values of k and then reaches a constant level. We then choose to balance the two needs of clusters made up of at least five elements, since the compatibility test requires a χ^2 evaluation, and of a high efficacy of the validation test (see Supplemental Material [53]).

V. VARIATIONS OF *K*-MEANS CLUSTERING

With K means, different initial conditions can lead to a different final structure. Hence, the algorithm can end up in a local minimum of its objective function. To avoid this issue, we consider three different strategies: (I),(II) replacing the random starting condition with two initialization algorithms, namely, (I) K means++ and (II) a preliminary run of hierarchical clustering, and (III) building on the same data set several cluster structures. (I) Once the user sets the number of clusters k, the first center is picked uniformly among the observed events. Then, for each observed element e, the distance $d(e)$ from the nearest of the picked centroids is evaluated. A new centroid is subsequently chosen randomly among the observed events, by assigning to each one a different weight given by $d(e)^2$. This procedure is iterated until all k cluster centers are initialized. Then, standard K-means clustering can be applied. (II) The user has to set the halting condition for hierarchical clustering. As discussed previously, in our case the process is interrupted when the fraction of outliers is smaller than a chosen threshold condition (≤ 0.01). The centroids of the final cluster structure obtained from hierarchical clustering are used as the starting condition for K means. (III) As said, when adopting K-means clustering, the final structure is not deterministic for a given data set. Hence, to reduce the variability of the final condition and thus avoid the algorithm getting stuck in a local minimum, the K-means method is run an odd number of times (for instance, 11), and majority voting is performed over the compatibility test results. Finally, the adoption of K means++ (I) and majority voting (III) can also be simultaneously combined.

VI. BENCHMARKING THE PROTOCOL

As a first step, we perform a detailed analysis to identify the most suitable among the mentioned clustering algorithms. More specifically, we proceed with the two following steps: (i) a *tuning stage* and (ii) a *cross-validation stage*. The figure of merit quantifying the capability of each test to perform correct decisions is the success percentage, i.e., the probability that two samples drawn from the same statistical population are labeled as compatible while two samples drawn from different probability distributions are recognized as incompatible.

(i) In the tuning stage, we look for the most effective clustering algorithm for our validation protocol. We are not yet applying it to validate boson-sampling data; rather, we are optimizing the set of hyperparameters that will define its operation (see Sec. VII and Supplemental Material [53]). Indeed, our protocol is based on unsupervised algorithms that, as such, do not need data with different labels to learn effective patterns.

To this aim, we apply all algorithms on numerically generated samples of output states, belonging to the collision-free subspace of $n = 3$ photons evolved through a fixed unitary transformation U with $m = 13$ modes. Hence, the dimension of the Hilbert space in this case is $\binom{13}{3} = 286$. Each algorithm is run several times, while varying the number of events within the tested samples. For each sample size, the hyperparameters proper of each technique are optimized. To evaluate the success percentages for each configuration of hyperparameters, we numerically generate 100 distinct data sets made of three samples: two of them are drawn from the boson-sampling distribution, while a third is drawn from the output probability distribution obtained when distinguishable particles are evolved with the same input state and unitary transformation U. We perform two compatibility tests for each data set: the first between two compatible samples and the second between two incompatible ones. The results of this analysis for samples with 500 output events are shown in Table II. We observe that the best success percentage is obtained for the K-means++ method with majority voting and employing the L_2 distance. The reason for which K means is outperforming bubble clustering lies in its learning capability. Indeed, due to its convergence properties through the iterations, K means gradually improves its insight into the internal structure that characterizes the data. This feature enables a better discrimination between compatible and incompatible samples (see Supplemental Material [53]).

(ii) In the cross-validation stage, we cross-validated the algorithm for random unitary transformations with a fixed size (ii.a) and for an increasing dimension of the Hilbert space (ii.b). As a first step (ii.a), we perform the test with $n = 3$ photons evolving through 20 Haar-random transformations with $m = 13$ modes. For each transformation, we perform 100 tests between compatible samples and 100 between incompatible ones, by fixing the number of clusters and trials to the values determined in stage (i). In Table III, we report the means and standard deviations of the success percentages for a sample size of 1000 events and compare the obtained values with the ones characterizing the bubble clustering method. We observe that the chosen approach, K means++ with majority voting and L_2 distance, indeed permits one to achieve better success percentages.

To extend the cross-validation to larger-dimensional Hilbert spaces (ii.b), we exploit an efficient algorithm to sample from the distribution with distinguishable photons [28] and a recent algorithm by Clifford and Clifford [58] to

TABLE II. Confusion matrix for different clustering techniques and fixed unitary evolution [tuning stage, step (i)]. Success percentages of the compatibility tests for all the different clustering techniques studied, i.e., bubble clustering, hierarchical clustering, and K-means clustering. The latter algorithm is investigated in its standard version and initialized by K means++ or a preliminary run of hierarchical clustering [57]. Then, majority voting is performed on the nondeterministic versions of K means. The reported success percentages are evaluated through numerical simulations by keeping the unitary evolution operator fixed. This choice is motivated by the need of tuning the different algorithms in order to subsequently classify new data sets.

			Output classification				
			1-norm		2-norm		
			Ind.	Dis.	Ind.	Dis.	
	(a) Bubble clustering		95 33	5 67	96 31	4 69	Ind. Dis.
	(b) Hierarchical clustering		1 2	99 98	8 5	92 95	Ind. Dis.
(c) K-means clustering	Uniformly distributed initialized centroids	Single trial	98 10	2 90	95 21	5 79	Ind. Dis.
		Majority voting	100 1	0 99	99 2	1 98	Ind. Dis.
	K means++ initialized centroids	Single trial	95 17	5 83	97 17	3 83	Ind. Dis.
		Majority voting	98 1	2 99	100 0	0 100	Ind. Dis.
	Hierarchical clustering initialized centroids		97 16	3 84	95 5	5 95	Ind. Dis.

TABLE III. Confusion matrix for bubble clustering and K means++ with majority voting random unitary evolution [cross-validation stage, step (ii.a)]. Success percentages of the compatibility test for bubble clustering and K means initialized with K means++ and majority voting. These percentages are evaluated through numerical simulations, by drawing 20 Haar-random unitary transformations, and by adopting the same hyperparameters obtained from stage (i) corresponding to the results in Table II.

		Output classification				
		1-norm		2-norm		
	Events	Ind.	Dis.	Ind.	Dis.	
Bubble	500	95.6 ± 2.8 69 ± 13	4.4 ± 2.8 31 ± 13	95.7 ± 1.7 75 ± 14	4.3 ± 1.7 25 ± 14	Ind. Dis.
	1000	95.9 ± 2.0 62 ± 30	4.1 ± 2.0 38 ± 30	93.1 ± 2.8 51 ± 23	6.9 ± 2.8 49 ± 23	Ind. Dis.
K-means++ m.v.	500	99.1 ± 1.2 45 ± 23	0.9 ± 1.2 55 ± 23	99.70 ± 0.57 66 ± 22	0.30 ± 0.57 34 ± 22	Ind. Dis.
	1000	98.7 ± 2.7 3.6 ± 6.4	1.3 ± 2.7 96.4 ± 6.4	96.2 ± 3.9 0.30 ± 0.73	3.8 ± 3.9 99.70 ± 0.73	Ind. Dis.

sample indistinguishable photons, with a much more efficient approach compared to the brute-force one. With this approach, we are able to test the efficacy of our protocol up to $n = 25$ photons in $m = 625$ modes (see Table IV). Note that, in this case, to validate a sample we require only 10^5 events, a negligible fraction (10^{-40}) of the total output states. In the case where $m \sim n^2$, while results are shown with fixed $N \sim 5 \times 10^4$ events for the sake of clarity and to avoid biases, in most instances $N \sim 10^4$ events are already sufficient. An aspect of our test that is worth noticing, as shown in Table III, is that the probability of error is lopsided, which is a feature that can be valuable for applications where falsely concluding that trustworthy boson samplers are unreliable is less desirable than the converse. Another crucial point is that we do not need to perform a tuning of the hyperparameters for each pair (n,m). We discuss this aspect in more detail in Sec. VII.

During the cross-validation stage, we also perform numerical simulations to verify whether the present approach is effective against other possible failure modes different from distinguishable particles, namely, the mean-field sampler [43] and a uniform sampler (see Sec. IV in Supplemental Material [53]). The former performs sampling from a suitable tailored single-particle distribution which

TABLE IV. Efficacy of K means++ for large-size boson sampling [cross-validation stage, step (ii.b)]. The algorithm is highly effective to discern quantum ($\mathcal{Q}$) boson samplers from classical ($\mathcal{C}$) and mean-field states ($\mathcal{MF}$), even for a large number of photons n and modes m and using very small sample sizes (N) as compared to the number of output combinations. For all probed (n, m), K means correctly identifies the nature of the $\mathcal{Q}$ sample, while, when tested with adversarial samples, it still correctly identifies all their instances after proper training. K means is initialized by K means++, with optimized hyperparameters and majority voting. Numerical samples of *bona fide* boson samplers are generated using the algorithm by Clifford and Clifford [58].

$\mathcal{C}, \mathcal{N} = 5 \times 10^4$	n							
m	3	4	5	6	7	8	9	10
9	✓	✓	✓	✓	✓	✓	✓	–
16	✓	✓	✓	✓	✓	✓	✓	✓
25	✓	✓	✓	✓	✓	✓	✓	✓
36	✓	✓	✓	✓	✓	✓	✓	✓
49	0.85	✓	✓	✓	✓	✓	✓	✓
64	0.80	✓	✓	✓	✓	✓	✓	✓
81	0.35	0.95	✓	✓	✓	✓	✓	✓
100	0.10	0.60	0.90	✓	✓	✓	✓	✓
	11	12	13	14	15	16	17	18
121	✓	✓	✓	✓	✓	✓	✓	✓
144	✓	✓	✓	✓	✓	✓	✓	✓
169	✓	✓	✓	✓	✓	✓	✓	✓
196	✓	✓	✓	✓	✓	✓	✓	✓
225	✓	✓	✓	✓	✓	✓	✓	✓

$\mathcal{MF}, \mathcal{N} = 5 \times 10^4$	n							
m	3	4	5	6	7	8	9	10
9	✓	✓	✓	✓	✓	✓	✓	–
16	✓	✓	✓	✓	✓	✓	✓	✓
25	✓	✓	✓	✓	✓	✓	✓	✓
36	0.90	✓	✓	✓	✓	✓	✓	✓
49	0.70	✓	✓	✓	✓	✓	✓	✓
64	0.50	0.95	✓	✓	✓	✓	✓	✓
81	0.30	0.80	0.95	0.95	✓	✓	✓	✓
100	0.05	0.50	0.80	0.95	✓	✓	✓	✓
	11	12	13	14	15	16	17	18
121	✓	✓	✓	✓	✓	✓	✓	✓
144	✓	✓	✓	✓	✓	✓	✓	✓
169	✓	✓	✓	✓	✓	✓	✓	✓
196	✓	✓	✓	✓	✓	✓	✓	✓
225	✓	✓	✓	✓	✓	✓	✓	✓

$\mathcal{N} = 10^5$	n
$m = 625$	25
$\mathcal{C}$	✓
$\mathcal{MF}$	✓

$\mathcal{N} = 2.5 \times 10^5$															
$m = 400$	3	4	5	6	7	8	9	10	11	12	13	14	15	…	20
$\mathcal{C}$	0	0	0	0.10	0.15	0.70	✓	✓	✓	✓	✓	✓	✓		✓
$\mathcal{MF}$	0	0	0.05	0.35	0.55	✓	✓	✓	✓	✓	✓	✓	✓		✓

✓	Correctly validates $\mathcal{Q}$, $\mathcal{C}$ and $\mathcal{MF}$ with no training, by partitioning in n clusters
p	Correctly validates $\mathcal{Q}$, instead validates $\mathcal{C}$ / $\mathcal{MF}$ with probability p without training
.	Correctly validates $\mathcal{Q}$, $\mathcal{C}$ and $\mathcal{MF}$ after training

reproduces the same features of multiphoton interference, while the latter performs sampling from a uniform distribution. We observe that the test is capable to distinguish between a *bona fide* boson sampler and a uniform or mean-field sampler. This capability highlights a striking feature of this algorithm, namely, the ability of our algorithm to generalize beyond the training set of distinguishable and indistinguishable samples used to learn the hyperparameters, into situations where the data come from approximations to the boson-sampling distribution that *prima facie* bear no resemblance to the initial training examples.

VII. SCALABLE TUNING OF THE HYPERPARAMETER

To increase the applicability of the compatibility test, we can choose suitable sets of hyperparameters for the embedded clustering algorithms. This preliminary stage, typical in machine learning, may require proper methods such as a grid search or randomized search and is, in general, very beneficial [51]. In the case of the clustering algorithms described in Sec. VI, one common hyperparameter is related to the minimum number of elements (sampled output events) assigned to any cluster. Specifically, bubble clustering, hierarchical clustering, and K means require one to set, respectively, the minimum cutoff radius, the maximum acceptable fraction of outliers (events belonging to clusters with less than N elements), and the number of clusters K. Also, these algorithms can be applied with different notions of distance, potentially beyond the L_1 and L_2 already discussed, to reflect different knowledge on the character of a system. In this section, we clarify how to best configure the protocol to operate in instances of large dimensionality, where no algorithm for classical simulations is available to probe its functioning.

In the following, let us then focus on the L_2 distance (see Sec. VI) and on the number of clusters K for K means, which we identify as the most effective technique for our purpose. We quantify the performance of the compatibility test with its accuracy, namely, the success probability in ruling out samples that are not compatible with quantum boson sampling. In particular, we study how the choice of K influences the test in two different scenarios: (i) when $m \sim n^2$ [Fig. 3(a)] and (ii) for the specific instance of

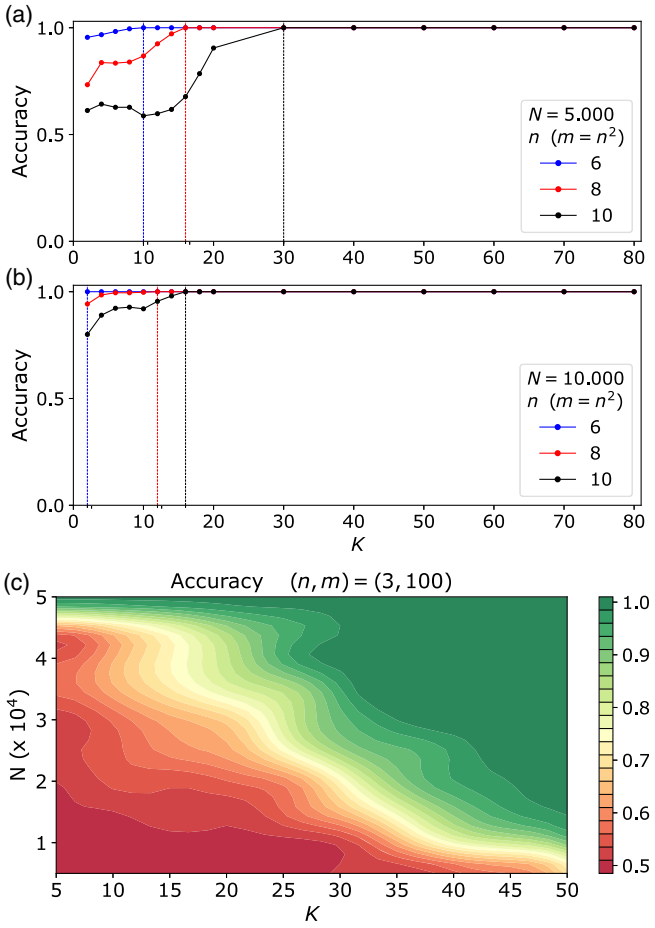

FIG. 3. Test accuracy versus number of clusters and sample size. The compatibility test based on K means improves its accuracy (ratio of correct assessments) with the number of samples N and clusters K. Hence, instead of fine-tuning K for each (n,m), we can directly set the hyperparameter to a large value, say, $(m/2) \leq K \leq m$, with negligible computational overhead [56]. This feature is investigated by numerically sampling $N = 5 \times 10^3$ (a) or $N = 10^4$ (b) photonic states with $n = 6, 8, 10$ photons in $m = n^2$ modes. Accuracies are estimated by applying the test to numerically simulated experiments with both indistinguishable and distinguishable photons from 200 Haar-random unitary transformations. Points are connected for the sake of clarity. (c) Contour plot for the accuracy in excluding classical boson sampling in the harder instance of $n = 3$ distinguishable photons in $m = 100$ modes, for different values of K and N. Colors describe the efficacy of the test, from red (poor) to green (perfect). All tests are performed considering a significance level of 5% for the χ^2 test and using K means.

$(n, m) = (3, 100)$, among the hardest ones presented in Table IV with $m \gg n$ [Fig. 3(b)]. Indeed, in the latter case, the probability of bunching is practically negligible, and the distributions with distinguishable and indistinguishable photons become much harder to discern for K means. From this analysis, we observe that the accuracy increases with the number of samples N, as expected, as well as with K. Indeed, the observation supports the intuition that a larger N provides more information to the algorithm to understand the spatial distribution in the Hilbert space, while a larger K allows one to probe it more finely. In particular, K should be sufficiently large to appreciate the detailed spatial dishomogeneities in the Hilbert space. Moreover, smaller values of K imply larger and more populated clusters that tend to average local fluctuations, so that less reliable evidence can be drawn from the compatibility test. Naturally, we also recall that K cannot be chosen overly large, since, in that case, each cluster would contain not enough points for the χ^2 test to make robust predictions. Thus, provided that the number of samples and clusters increases accordingly, the protocol is effective even in the nonfavorable condition $(n, m) = (3, 100)$ [see Table IV and Fig. 3(b)]. In the regime where $m \sim n$ or $m \sim n^2$ [Fig. 3(a)], the algorithm is instead successful for almost all choices of K when $N \sim 10^4$–10^5, again in accordance with Table IV. As a general rule of thumb, which proves effective in all combinations (n,m) investigated, we set $(m/2) \leq K \leq m$ to satisfy the need for a bounded-above value that grows with the size of the problem. Anyway, the ultimate relevance of this specific choice can be further relaxed by collecting a larger number of samples. In this sense, the analysis reported in Fig. 3 removes the burden of a fine-tuning at low dimensions, thus extending the applicability of the protocol. Moreover, the whole approach gains much also in terms of simplicity, a feature that can prove beneficial for practical applications besides the assessment of quantum supremacy.

VIII. EXPERIMENTAL RESULTS

Through the experimental apparatus shown in Fig. 4(a), we collect samples corresponding to the boson-sampling distribution with indistinguishable and distinguishable particles. The degree of distinguishability between the input photons is adjusted by modifying their relative arrival times through delay lines (see Supplemental Material [53]). The unitary evolution is implemented by an integrated photonic chip realized exploiting the 3D-geometry capability of femtosecond laser writing [59] and performs the same transformation U employed for the numerical results in Table II. We then perform the same compatibility tests described previously on experimental data sets with different sizes, by using two methods: K means++ with majority voting and bubble clustering, both with 2-norm distance. The results are shown in Fig. 4(a), for the case of incompatible samples. This result implies that the reported percentages represent the capability of the test to recognize a boson sampler fed with distinguishable photon inputs. Reshuffling of the experimental data is used to have a sufficient number of samples to evaluate the success percentages (see Supplemental Material [53]). Hence, the tests are performed on samples drawn randomly from the experimental data.

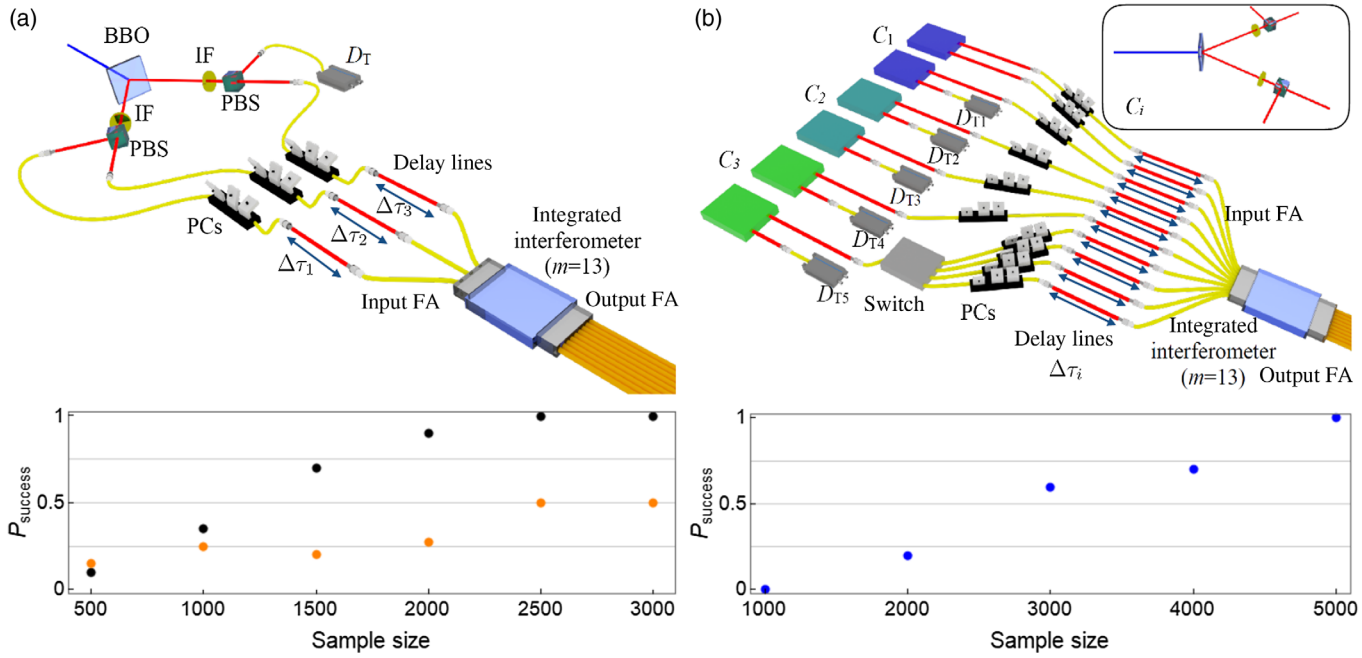

FIG. 4. Experimental validation of boson-sampling experiments. Experimental setups for an $n = 3$ standard (a) and scatter-shot (b) boson-sampling experiments in an integrated $m = 13$ interferometer (see Supplemental Material [53]). Bottom insets: The corresponding success probabilities of the compatibility test between inputs with indistinguishable and distinguishable photons for different sample sizes. (a) Black and orange dots refer, respectively, to K means++ with majority voting and bubble clustering. The discrepancy from the numerical results in Table III is due to the nonideal indistinguishability of the injected photons [60]. (b) Input states are generated probabilistically by six independent parametric down-conversion sources (represented as boxes) in three different BBO crystals C_i (see the inset). Experimental data correspond to eight different inputs. The number of events for all input states randomly varies for each sample size drawn from the complete data set. The clustering algorithm is K means, initialized by K means++, with majority voting. In both panels: BBO, beta barium borate crystal; IF, interferential filter; PBS, polarizing beam splitter; PC, polarization controller; FA, fiber array. Tests are performed with a significance level of 5% and using $d = L_2$.

IX. GENERALIZATION FOR SCATTER-SHOT BOSON SAMPLING

The scatter-shot version of boson sampling [19,20] is implemented through the setup in Fig. 4(b). Six independent parametric down-conversion photon pair sources are connected to different input modes of the 13-mode integrated interferometer. In this case, two input modes (6 and 8) are always fed with a single photon. The third photon is injected probabilistically into a variable mode, and the input is identified by the detection of the twin photon at trigger detector T_i. We consider a generalization of the proposed algorithm to be applied for scatter-shot boson sampling. In this variable-input scenario, a boson sampler to be validated provides N samples that correspond to N different inputs of the unitary transformation, that is, N Fock states Φ_i with $i \in \{1, n\}$. Hence, our validation algorithm in its standard version needs to perform N separate compatibility tests. Indeed, it brings N distinct chi-square variables χ_i^2, where the ith variable quantifies the agreement between the distribution of the data belonging to the input Φ_i and the distribution of a sample drawn by a trusted boson sampler with the same input state. Hence, each input state is tested separately.

In order to extract only one quantity to tell whether the full data set is validated or not, for all inputs, a new variable can be defined as $\tilde{\chi}^2 = \sum_{i=1}^{N} \chi_i^2$. This variable is a chi-square one with $\nu = \sum_{i=1}^{N} \nu_i$ d.o.f., provided that the χ_i^2 are independent. We perform this generalized test on the experimental data by adopting the same clustering technique previously discussed in the single-input case.

X. EXPERIMENTAL RESULTS FOR SCATTER-SHOT BOSON SAMPLING

We collect output samples given by eight different inputs both with indistinguishable photons and with distinguishable ones. Through the evaluation of the new variable $\tilde{\chi}^2$, the algorithm is able to distinguish between a trustworthy scatter-shot boson sampler and a fake one at the significance level of 5%, using a total number of observed events up to 5000 events (over all inputs), as shown in Fig. 4(b). The standard version of the test, validating each input separately, requires samples of 2000 events per input to reach a success percentage $\geq 80\%$, that is, an overall amount of 16 000 events. Hence, the generalized version of the test permits one to significantly reduce the amount of necessary resources to validate scatter-shot boson-sampling experiments.

XI. STRUCTURE OF THE PROBABILITY DISTRIBUTIONS

Our previous discussion conclusively shows that, at least for the values of (n, m) considered, K-means clustering algorithms are highly effective at discriminating boson samplers that use distinguishable photons versus those with indistinguishable ones. Here, we provide further analysis that shows why our approach is effective at this task and sheds light on how future tests could be devised to characterize faulty boson samplers. We address this aspect by providing numerical evidence to explain the physical mechanism behind the correct functioning of our validation test.

The clustering techniques that form the basis of our pattern-recognition methodology rely on aggregating the experimental data according to the distance between the output states. The key observation is that the number of events necessary to effectively discriminate the samples is dramatically lower than the number of available output combinations. In particular, such a fraction drops fast to smaller values by increasing the system size (n, m). For instance, 10^4 events correspond to 0.08 of the Hilbert space dimension for (4, 40), to 10^{-10} for (10, 100), and to only 10^{-30} for (20, 400). Accordingly, the output sample from the device mostly consists of output states occurring with no repetition. Hence, only the configurations presenting higher probability effectively contribute to the validation test.

For the sake of clarity, let us focus on the discrimination between indistinguishable and distinguishable particles. We leave for subsequent work the task of explaining why other alternative models, such as mean-field states, are also noticed by our approach. More specifically, we analyze the structure of the outcome distributions for the two cases. Since data clustering is performed according to the distance between states, the method can be effective if (i) the distributions of the output states exhibit an internal structure and (ii) correlations between distributions with different particle types are low.

As a first step towards this goal, we compute the probability distributions with $n = 4$ indistinguishable photons (P_j) and distinguishable particles (Q_j), for a fixed unitary transformation U with $m = 40$ modes. Figure 5(a) reports the two distributions sorted according to the following procedure, in order to highlight their different internal structure. The distribution with indistinguishable photons is sorted in decreasing order starting from the highest probability, while the distribution with distinguishable particles is sorted by following the same order adopted for the indistinguishable case. More specifically, the first element is the value of Q_j for the output state corresponding to the highest value of P_j, the second element corresponds to the state with the second-highest value of P_j, and analogously for all other terms. We observe a small correlation between the P_j and Q_j distributions. To quantify this feature, we compute two different statistical coefficients (the Pearson r and the Spearman's rank ρ ones), that are employed to evaluate the presence of linear or generic correlations between two random variables. In particular, we find that the Pearson correlation coefficient is $r \sim 0.56$, while the Spearman's rank coefficient is $\rho \sim 0.55$, which suggests that the two distributions have different supports over the outcome distributions. The same analysis is performed for $n = 5$ and $m = 50$, showing that a similar behavior is obtained for increasing size [see Fig. 5(b)]. By averaging over $M' = 100$ different unitaries, the correlation coefficients are $r \sim 0.62 \pm 0.03$ and $\rho \sim 0.64 \pm 0.04$ (1 standard deviation) for $n = 4$ and $m = 40$ and $r \sim 0.57 \pm 0.03$ and $\rho \sim 0.62 \pm 0.04$ (1 standard deviation) for $n = 5$ and $m = 50$. These results show that the low values of the correlations between P_j and Q_j do not depend on the specific transformation U and that this behavior is maintained for larger size systems. Similar conclusions are observed in the cumulative distributions [see Fig. 5(c)], where the distinguishable case is sorted by following the same order as the indistinguishable one. We observe that, for the cumulative probability for distinguishable bosons to reach the same value attained for indistinguishable bosons, a significantly larger portion of the Hilbert space has to be included. For instance, when $n = 4$ and $m = 40$, 50% of the overall probability is achieved by using approximately 13% of the overall number of outputs for indistinguishable photons, while approximately 32% are necessary for the distinguishable case (by following the above-mentioned ordering procedure). Similar numbers are obtained for larger dimensionalities (approximately 11% and approximately 30%, respectively, when $n = 5$ and $m = 50$).

The second crucial aspect of our method is related to the localization of outcomes with the highest probabilities. More specifically, this approach can be effective in constructing useful cluster structures if the most probable states are surrounded by other states with high probability. In this way, when a number of events much lower than the number of combinations is collected, the outcomes actually occurring in the data sample present lower distance values, thus justifying the application of a clustering procedure. An intuition for this feature is provided in Figs. 5(d) and 5(e), which show that centroids tend to locate in the positions of the m-dimensional vector space corresponding to the output modes with the highest probability, averaged over the input modes.

We further probe how these correlations become visible through a clustering method by performing numerical simulations that randomly vary the unitary transformation U for $(n = 4,\ m = 40)$ and $(n = 5,\ m = 50)$. For each sampled transformation U, we calculate the probabilities P_j and Q_j for both cases (indistinguishable and distinguishable photons) and then sort the distribution P_j in decreasing order. Let us call J the outcome with the highest P_j value, which is to say $J = \mathrm{argmax}(P_j)$. Let us for simplicity fix the distance to be the L_1-norm (analogous results are obtained

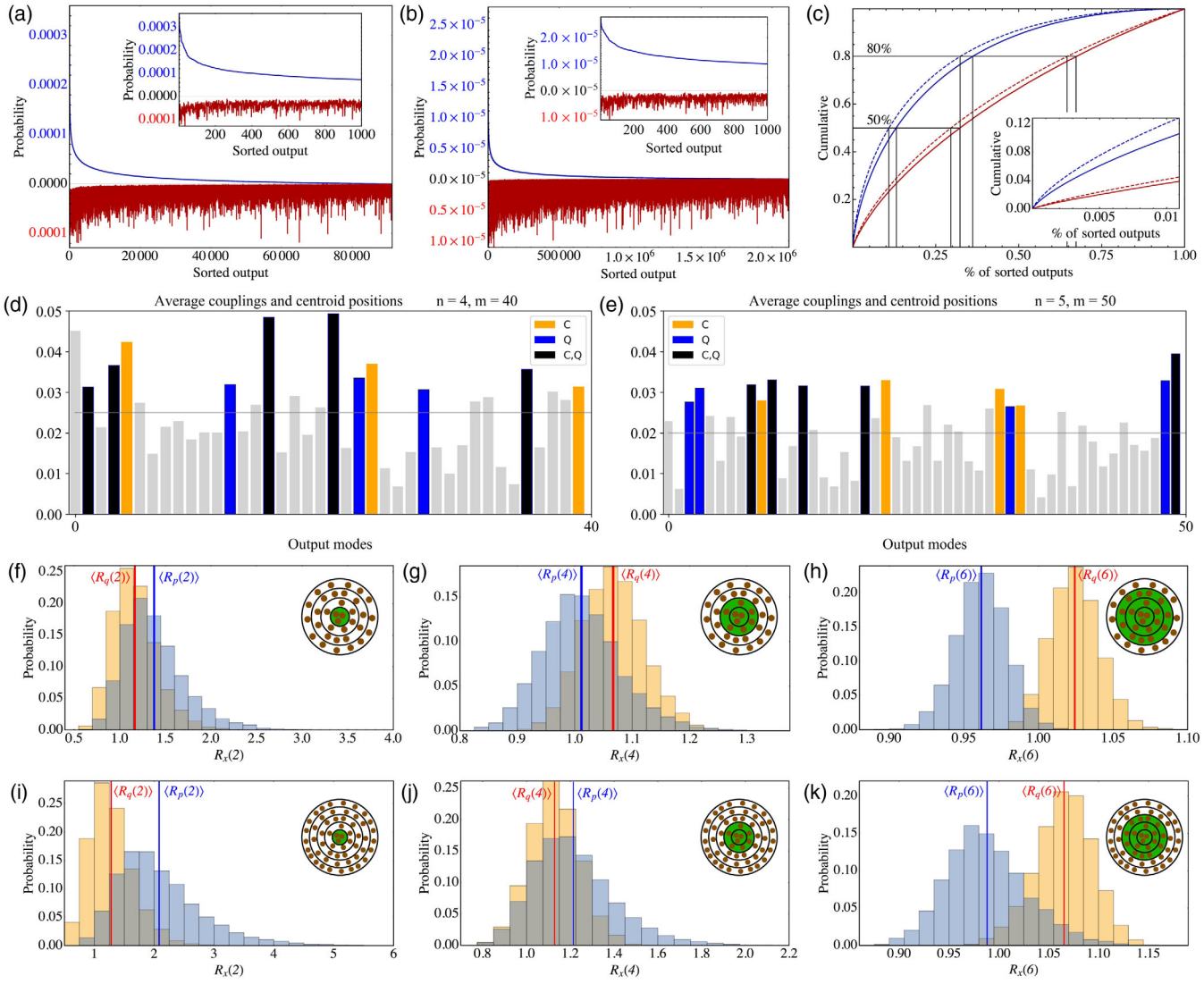

FIG. 5. Analysis of the structure of the distributions. (a),(b) Probability distributions for a fixed unitary U in the case of indistinguishable (blue) and distinguishable (red) photons. (a) $n = 4$, $m = 40$ and (b) $n = 5$, $m = 50$. The distributions are sorted by following the same ordering so as to have decreasing probability values for the indistinguishable distribution P_i. Inset: Enlargement corresponding to the 1000 most probable output states. (c) Cumulative distributions for indistinguishable (blue) and distinguishable (red) photons, by following the same ordering as (a),(b). Solid lines, $n = 4$, $m = 40$. Dashed lines, $n = 5$, $m = 50$. Black lines highlight the levels corresponding to 50% and 80% of the overall probability, which require approximately twice the number of output states in the distinguishable case. Inset: Enlargement corresponding to the 0.01% most probable output states. (d),(e) In the m-dimensional vector space where K means is performed on quantum (Q) and classical (C) samples, centroids tend to be located where the output modes have a high probability averaged over the input modes. This property is signaled by a vector with one coordinate, out of m for a Fock state, significantly higher than the others. Zeroing the small ones out yields a plot analogous to the one shown here for ten clusters with (d) $n = 4$, $m = 40$ and (e) $n = 5$, $m = 50$. (f)–(k) Histograms of the ratios $R_p(k)$ (cyan) and $R_q(k)$ (orange) between the overall probability included within a sphere of L_1-norm $\leq k$. (f),(h) $n = 4, m = 40$. (i)–(k) $n = 5, m = 50$. Vertical lines correspond to the averages $\langle R_x(k)\rangle$, with $x = p$ (blue) and $x = q$ (red). (f),(i) $k = 2$, (g),(j) $k = 4$, and (h),(k) $k = 6$. Insets: Schematic view of the spheres at a distance $\leq k$, represented by concentric circles, where states are represented by brown points.

for the L_2-norm). Note that the L_1-norm defined in the main text has only N possible nontrivial values $k = 2s$, with $s = 1, \ldots, n$. We then estimate the overall probability $P(k) = \sum_{j:\|j-J\|_1 \leq k} P_j$, where $P(k)$ is the probability included in a sphere with a distance $\leq k$ computed using the L_1 norm. The same calculation is performed for the distinguishable particle case $Q(k) = \sum_{j:\|j-J\|_1 \leq k} Q_j$, by using the same outcome value J as a reference.

We study the ratio $R_p(k) = P(k)/Q(k)$ between the two probabilities, that can be thought of as a likelihood ratio test, wherein $R_p(k) > 1$ implies that the evidence is in favor of indistinguishable particles and, conversely, $R_p(k) < 1$

suggests that the particles are distinguishable. Such a comparison is then performed for $M'' = 100$ different unitary matrices U and by using as reference outcome J the $M_{\max} = 100$ highest-probability outcomes for each U. The results are reported in Figs. 5(d)–5(f) for $n = 4, m = 40$ with $k = 2, 4, 6$ (being $k = 8$ a trivial one, which includes all output states given four-photon input states). The analysis is also repeated in the opposite case, where the data are sorted according to the distinguishable particle distribution Q_j and $R_q = Q(k)/P(k)$. We observe that $R_p(2)$ has an average of $\langle R_p(2)\rangle \sim 1.4$ and that $P[R_p(2) > 1] \sim 0.904$. For increasing values of k, $\langle R_p(k)\rangle$ converges to unity, since a progressively larger portion of the Hilbert space is included, thus converging to $R_p(k = 8) = 1$ (respectively, approximately 0.16% for $k = 2$, approximately 4.3% for $k = 4$, and approximately 35.5% for $k = 6$). Similar results are obtained also for $n = 5$ and $m = 50$ [see Figs. 5(g)–5(j)], where $\langle R_p(2)\rangle \sim 2.08$, and that $P[R_p(2) > 1] \sim 0.986$. This behavior for $R_p(k)$ and $R_q(k)$ highlights a hidden correlation within the output state distributions, where outcomes with higher probabilities tend to be more strongly localized at low L_1 distance from the reference outcome for indistinguishable bosons than those drawn from a distribution over distinguishable particles. This behavior is why basing an algorithm on this feature is effective for diagnosing a faulty boson sampler that uses distinguishable photons.

The same considerations can be obtained also from a different perspective. Indeed, it has been recently shown [36] that information on particle statistics from a multiparticle experiment can be retrieved by low-order correlation measurements of $C_{ij} = \langle n_i n_j\rangle - \langle n_i\rangle\langle n_j\rangle$, where n_i is the number operator. Correlations between the states of the output distribution, originating from the submatrix of U that determines all output probabilities, correspond to correlations between the output modes. Such correlations are different depending on the particle statistics (indistinguishable or distinguishable particles) due to interference effects and can thus be exploited to identify the particle type given an output data sample. More specifically, a difference between particle types is observed in the moments of the C_{ij} set, thus highlighting different structures in the output distributions. As previously discussed, such different structures can be detected by clustering approaches.

In summary, all these analyses show that boson-sampling distributions with indistinguishable and distinguishable particles present an internal structure that can be caught by the clustering procedure at the basis of our validation method, thus rendering our method effective to discriminate between the two hypotheses.

XII. DISCUSSION

In this article, we show that pattern-recognition techniques can be exploited to identify pathologies in boson-sampling experiments. The main feature of the devised approach relies on the absence of any permanent evaluation, thus not requiring the calculation of hard-to-compute quantities during the process. The efficacy of these pattern-recognition techniques relies on the presence of marked correlations in the output distributions that are related to the localization of the outcomes with the highest probabilities and that depend on the particle type. Additionally, the absence of any assumptions on the system under study allows one to apply the compatibility test to a much broader class of multiparticle states: One just needs a trusted sample from an arbitrary class to check whether or not a different sample is compatible with it.

This approach can also be adopted in larger Hilbert spaces with arguably no need for a fine-tuning of clustering hyperparameters, which makes it a promising approach for identifying flaws in the next generation of boson samplers. We can further adopt the test as part of the validation toolbox at the boundaries of quantum supremacy, where classical and quantum sampling take approximately the same time, and it is still possible to numerically generate the trusted samples.

We also envisage that other protocols, based on more sophisticated machine-learning methods, might in the future provide even more effective solutions to this aim. Moreover, our experimental demonstration shows that it is possible to successfully test boson sampling even in lossy scenarios, which have already been shown to maintain the same computational hardness of the original problem [61]. Looking forward, it is our hope that when building data-driven (rather than first-principles) models for error, cross-validation will be used to report the performance of such algorithms. For example, our method has 100% classification accuracy for the training data but has roughly 95% accuracy in the test data. Had we reported only the performance of the algorithm on the training data, it would have provided a misleading picture of the method's performance for larger boson-sampling experiments. For this reason, it is important that, if we are to use the tools of machine learning to help validate quantum devices, then we should also follow the lessons of machine learning when reporting our results.

While our work shows that machine learning can be used to provide evidence that a boson sampler is faulty, it does not provide a definitive test. Furthermore, even if a boson sampler passes such tests, it need not also be a valid boson sampler, which means that, while machine learning is a valuable tool to help build confidence in boson samplers, it does not solve the validation problem in and of itself. Finding ways to clearly state the assumptions under which such machine-learning approaches validate a boson sampler, and the *a posteriori* probability with which it is found to be valid, remains an open problem.

Finally, although our work is focused on the validation of boson samplers, it is important to note that the lessons learned from this task are more generally applicable. Unsupervised methods, such as clustering, can be used to find patterns in high-dimensional data that allow simple

algorithms to learn facts about complex quantum systems that humans can easily miss. As a simple example, we show that the centroids' positions are correlated to the modes where the single-particle probability is higher on average. By continuing to incorporate ideas from computer vision into our verification and validation toolbox, we may not only develop the toolbox necessary to provide a convincing counterexample to the extended Church-Turing thesis, but also provide the means to debug the first generation of fault-tolerant quantum computers.

ACKNOWLEDGMENTS

This work was supported by the ERC-Starting Grant 3D-QUEST (3D-Quantum Integrated Optical Simulation; Grant Agreement No. 307783), by the H2020-FETPROACT-2014 Grant QUCHIP (Quantum Simulation on a Photonic Chip; Grant Agreement No. 641039), and by the European Research Council (ERC) Advanced Grant CAPABLE (Composite integrated photonic platform by femtosecond laser micromachining, Grant Agreement No. 742745).

[1] P. W. Shor, *Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer*, SIAM J. Sci. Stat. Comput. **26**, 1484 (1997).

[2] M. Schuld, I. Sinayskiy, and F. Petruccione, *An Introduction to Quantum Machine Learning*, Contemp. Phys. **56**, 172 (2015).

[3] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and J. Biamonte, *Quantum Machine Learning*, Nature (London) **549**, 195 (2015).

[4] R. Feynman, *Simulating Physics with Computers*, J. Theor. Phys. **21**, 467 (1982).

[5] S. Lloyd, *Universal Quantum Simulators*, Science **273**, 1073 (1996).

[6] M. J. Bremner, A. Montanaro, and D. J. Shepherd, *Average-Case Complexity versus Approximate Simulation of Commuting Quantum Computations*, Phys. Rev. Lett. **117**, 080501 (2016).

[7] M. J. Bremner, A. Montanaro, and D. Shepherd, *Achieving Quantum Supremacy with Sparse and Noisy Commuting Quantum Computations*, Quantum **1**, 8 (2017).

[8] X. Gao, S.-T. Wang, and L.-M. Duan, *Quantum Supremacy for Simulating a Translation-Invariant Ising Spin Model*, Phys. Rev. Lett. **118**, 040502 (2017).

[9] J. Bermejo-Vega, D. Hangleiter, M. Schwarz, R. Raussendorf, and J. Eisert, *Architectures for Quantum Simulation Showing a Quantum Speedup*, Phys. Rev. X **8**, 021010 (2018).

[10] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, *Characterizing Quantum Supremacy in Near-Term Devices*, Nat. Phys. **14**, 595 (2018).

[11] S. Aaronson and A. Arkhipov, *The computational complexity of linear optics*, in *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, San Jose, California (Association for Computing Machinery, New York, 2011), pp. 333–342.

[12] M. A. Broome, A. Fedrizzi, S. Rahimi-Keshari, J. Dove, S. Aaronson, T. C. Ralph, and A. G. White, *Photonic Boson Sampling in a Tunable Circuit*, Science **339**, 794 (2013).

[13] J. B. Spring, B. J. Metcalf, P. C. Humphreys, W. S. Kolthammer, X.-M. Jin, M. Barbieri, A. Datta, N. Thomas-Peter, N. K. Langford, D. Kundys, J. C. Gates, B. J. Smith, P. G. R. Smith, and I. A. Walmsley, *Boson Sampling on a Photonic Chip*, Science **339**, 798 (2013).

[14] M. Tillmann, B. Dakic, R. Heilmann, S. Nolte, A. Szameit, and P. Walther, *Experimental Boson Sampling*, Nat. Photonics **7**, 540 (2013).

[15] A. Crespi, R. Osellame, R. Ramponi, D. J. Brod, E. F. Galvao, N. Spagnolo, C. Vitelli, E. Maiorino, P. Mataloni, and F. Sciarrino, *Integrated Multimode Interferometers with Arbitrary Designs for Photonic Boson Sampling*, Nat. Photonics **7**, 545 (2013).

[16] N. Spagnolo, C. Vitelli, L. Sansoni, E. Maiorino, P. Mataloni, F. Sciarrino, D. J. Brod, E. F. Galvao, A. Crespi, R. Ramponi, and R. Osellame, *General Rules for Bosonic Bunching in Multimode Interferometers*, Phys. Rev. Lett. **111**, 130503 (2013).

[17] N. Spagnolo, C. Vitelli, M. Bentivegna, D. J. Brod, A. Crespi, F. Flamini, S. Giacomini, G. Milani, R. Ramponi, P. Mataloni, R. Osellame, E. F. Galvao, and F. Sciarrino, *Experimental Validation of Photonic Boson Sampling*, Nat. Photonics **8**, 615 (2014).

[18] J. Carolan, J. D. A. Meinecke, P. J. Shadbolt, N. J. Russell, N. Ismail, K. Worhoff, T. Rudolph, M. G. Thompson, J. L. O'Brien, J. C. F. Matthews, and A. Laing, *On the Experimental Verification of Quantum Complexity in Linear Optics*, Nat. Photonics **8**, 621 (2014).

[19] A. P. Lund, A. Laing, S. Rahimi-Keshari, T. Rudolph, J. L. OBrien, and T. C. Ralph, *Boson Sampling from a Gaussian State*, Phys. Rev. Lett. **113**, 100502 (2014).

[20] M. Bentivegna, N. Spagnolo, C. Vitelli, F. Flamini, N. Viggianiello, L. Latmiral, P. Mataloni, D. J. Brod, E. F. Galvao, A. Crespi, R. Ramponi, R. Osellame, and F. Sciarrino, *Experimental Scattershot Boson Sampling*, Sci. Adv. **1**, e1400255 (2015).

[21] J. Carolan, C. Harrold, C. Sparrow, E. Martin-Lopez, N. J. Russell, J. W. Silverstone, P. J. Shadbolt, N. Matsuda, M. Oguma, M. Itoh, G. D. Marshall, M. G. Thompson, J. C. F. Matthews, T. Hashimoto, J. L. O'Brien, and A. Laing, *Universal Linear Optics*, Science **349**, 711 (2015).

[22] J. C. Loredo, M. A. Broome, P. Hilaire, O. Gazzano, I. Sagnes, A. Lemaitre, M. P. Almeida, P. Senellart, and A. G. White, *Boson Sampling with Single-Photon Fock States from a Bright Solid-State Source*, Phys. Rev. Lett. **118**, 130503 (2017).

[23] Y. He, X. Ding, Z.-E. Su, H.-L. Huang, J. Qin, C. Wang, S. Unsleber, C. Chen, H. Wang, Y.-M. He, X.-L. Wang, W.-J. Zhang, S.-J. Chen, C. Schneider, M. Kamp, L.-X. You, Z. Wang, S. Hofling, C.-Y. Lu, and J.-W. Pan, *Time-Bin-Encoded Boson Sampling with a Single-Photon Device*, Phys. Rev. Lett. **118**, 190501 (2017).

[24] H. Wang, Y. He, Y.-H. Li, Z.-E. Su, B. Li, H.-L. Huang, X. Ding, M.-C. Chen, C. Liu, J. Qin, J.-P. Li, Y.-M. He, C. Schneider, M. Kamp, C.-Z. Peng, S. Hofling, C.-Y. Lu, and J.-W. Pan, *High-Efficiency Multiphoton Boson Sampling*, Nat. Photonics **11**, 361 (2017).

[25] H. Wang, W. Li, X. Jiang, Y.-M. He, Y.-H. Li, X. Ding, M.-C. Chen, J. Qin, C.-Z. Peng, C. Schneider, M. Kamp, W.-J. Zhang, H. Li, L.-X. You, Z. Wang, J. P. Dowling, S. Hofling, C.-Y. Lu, and J.-W. Pan, *Toward Scalable Boson Sampling with Photon Loss*, Phys. Rev. Lett. **120**, 230502 (2018).

[26] C. Shen, Z. Zhang, and L. M. Duan, *Scalable Implementation of Boson Sampling with Trapped Ions*, Phys. Rev. Lett. **112**, 050504 (2014).

[27] C. Gogolin, M. Kliesch, L. Aolita, and J. Eisert, *Boson-Sampling in the Light of Sample Complexity*, arXiv: 1306.3995.

[28] S. Aaronson and A. Arkhipov, *Boson Sampling Is Far from Uniform*, Quantum Inf. Comput. **14**, 1383 (2014).

[29] N. Wiebe, *Using Quantum Computing to Learn Physics*, Bull. EATCS **112**, 1 (2014).

[30] L. Aolita, C. Gogolin, M. Kliesch, and J. Eisert, *Reliable Quantum Certification of Photonic State Preparations*, Nat. Commun. **6**, 8498 (2015).

[31] C. K. Hong, Z. Y. Ou, and L. Mandel, *Measurement of Subpicosecond Time Intervals between Two Photons by Interference*, Phys. Rev. Lett. **59**, 2044 (1987).

[32] A. J. Menssen, A. E. Jones, B. J. Metcalf, M. C. Tichy, S. Barz, W. S. Kolthammer, and I. A. Walmsley, *Distinguishability and Many-Particle Interference*, Phys. Rev. Lett. **118**, 153603 (2017).

[33] M. Bentivegna, N. Spagnolo, C. Vitelli, D. J. Brod, A. Crespi, F. Flamini, R. Ramponi, P. Mataloni, R. Osellame, E. F. Galvao, and F. Sciarrino, *Bayesian Approach to Boson Sampling Validation*, Int. J. Quantum. Inform. **12**, 1560028 (2014).

[34] N. Viggianiello, F. Flamini, M. Bentivegna, N. Spagnolo, A. Crespi, D. J. Brod, E. F. Galvao, R. Osellame, and F. Sciarrino, *Optimal Photonic Indistinguishability Tests in Multimode Networks*, Sci. Bull. **63**, 1470 (2018).

[35] V. S. Shchesnovich, *Universality of Generalized Bunching and Efficient Assessment of Boson Sampling*, Phys. Rev. Lett. **116**, 123601 (2016).

[36] M. Walschaers, J. Kuipers, J.-D. Urbina, K. Mayer, M. C. Tichy, K. Richter, and A. Buchleitner, *Statistical Benchmark for Boson Sampling*, New J. Phys. **18**, 032001 (2016).

[37] M. Bentivegna, N. Spagnolo, and F. Sciarrino, *Is My Boson Sampler Working?*, New J. Phys. **18**, 041001 (2016).

[38] K. Liu, A. P. Lund, Y.-J. Gu, and T. C. Ralph, *A Certification Scheme for the Boson Sampler*, J. Opt. Soc. Am. B **33**, 1835 (2016).

[39] T. Giordani, F. Flamini, M. Pompili, N. Viggianiello, N. Spagnolo, A. Crespi, R. Osellame, N. Wiebe, M. Walschaers, A. Buchleitner, and F. Sciarrino, *Experimental Statistical Signature of Many-Body Quantum Interference*, Nat. Photonics **12**, 173 (2018).

[40] C. Dittel, R. Keil, and G. Weihs, *Many-Body Quantum Interference on Hypercubes*, Quantum Sci. Technol. **2**, 1 (2017).

[41] C. Dittel, G. Dufour, M. Walschaers, G. Weihs, A. Buchleitner, and R. Keil, *Totally Destructive Many-Particle Interference*, Phys. Rev. Lett. **120**, 240404 (2018).

[42] N. Viggianiello, F. Flamini, L. Innocenti, D. Cozzolino, M. Bentivegna, N. Spagnolo, A. Crespi, D. J. Brod, E. F. Galvao, R. Osellame, and F. Sciarrino, *Experimental Generalized Quantum Suppression Law in Sylvester Interferometers*, New J. Phys. **20**, 033017 (2018).

[43] M. C. Tichy, K. Mayer, A. Buchleitner, and K. Molmer, *Stringent and Efficient Assessment of Boson-Sampling Devices*, Phys. Rev. Lett. **113**, 020502 (2014).

[44] A. Crespi, *Suppression law for Multiparticle Interference in Sylvester Interferometers*, Phys. Rev. A **91**, 013811 (2015).

[45] A. Crespi, R. Osellame, R. Ramponi, M. Bentivegna, F. Flamini, N. Spagnolo, N. Viggianiello, L. Innocenti, P. Mataloni, and F. Sciarrino, *Suppression Law of Quantum States in a 3D Photonic Fast Fourier Transform Chip*, Nat. Commun. **7**, 10469 (2016).

[46] S. T. Wang and L.-M. Duan, *Certification of Boson Sampling Devices with Coarse-Grained Measurements*, arXiv: 1601.02627.

[47] L. G. Valiant, *The Complexity of Computing the Permanent*, Theor. Comput. Sci. **8**, 189 (1979).

[48] S. Scheel, *Permanent in Linear Optical Networks*, arXiv: quant-ph/0406127.

[49] P. Simon, *Too Big to Ignore: The Business Case for Big Data* (Wiley, New York, 2013).

[50] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

[51] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT, Cambridge, MA, 2012).

[52] L. Rokach and O. Maimon, *Data Mining and Knowledge Discovery Handbook* (Springer, New York, 2005), Chap. Clustering Methods.

[53] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.9.011013 for more details on the validation algorithm, on the data analysis, and on the experimental apparatus.

[54] J. MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (University of California, Berkeley, 1967), Vol. 1, pp. 281–297.

[55] E. W. Forgy, *Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications*, Biometrics **21**, 768 (1965).

[56] S. P. Lloyd, *Least squares quantization in PCM*, IEEE Trans. Inf. Theory **28**, 129 (1982).

[57] Though K means is not guaranteed to converge for metrics other than L_2, L_1 proved equally effective in our protocol. In this case, it is also possible to use K medians, since the median is the best L_1 estimator just like the mean for L_2.

[58] P. Clifford and R. Clifford, *The Classical Complexity of Boson Sampling*, in *SODA '18: Proceedings of the 29th ACM-SIAM Symposium on Discrete Algorithms* (SIAM (Society for Industrial and Applied Mathematics), Philadelphia, Pennsylvania, USA, 2018), pp. 146–155.

[59] R. Gattass and E. Mazur, *Femtosecond Laser Micromachining in Transparent Materials*, Nat. Photonics **2**, 219 (2008).

[60] N. Spagnolo, C. Vitelli, L. Aparo, P. Mataloni, F. Sciarrino, A. Crespi, R. Osellame, and R. Ramponi, *Three-Photon Bosonic Coalescence in an Integrated Tritter*, Nat. Commun. **4**, 1606 (2013).

[61] S. Aaronson and D. J. Brod, *Boson Sampling with Lost Photons*, Phys. Rev. A **93**, 012335 (2016).