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, WA, United States*

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. To address this problem, 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 train 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 scattershot experiments. The effectiveness of such 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.

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 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 schemes have been proposed in the last few years [6–11] 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. A significant step in this direction has been achieved in particular by Aaronson and Arkhipov [6] with the formal definition of a dedicated task known as Boson Sampling. This 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 [6]. 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 [12–23] and also alternative approaches have been proposed, for example exploiting trapped ions [24] or applying random gates in superconducting qubits [8].

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 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 [25–27] and it represents a critical point for all the above-mentioned platforms seeking a first demonstration of quantum supremacy. A first approach to ensure quantum interference could involve testing pairwise mutual
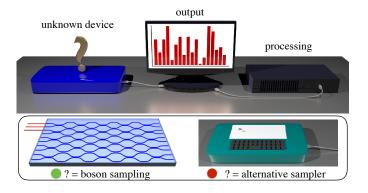

FIG. 1. **Validation of Boson Sampling experiments.** An agent has to discriminate whether a finite sample obtained from an unknown device has been generated by a quantum device implementing the Boson Sampling problem or by an alternative sampler.

---

* fabio.sciarrino@uniroma1.it

indistinguishability by two-photon Hong-Ou-Mandel experiments [28], however such method fails to completely characterize multiphoton interference [29]. While techniques exist that use likelihood ratios or cross-entropy to validate [17, 30], they work only for small systems. Other existing techniques exploit statistical properties of bosonic states [18, 31–33] or symmetries of certain Boson Samplers [20, 34–36], however these methods are much more limited in scope.

Our approach in this article is different. Rather than building our tests upon our limited physical intuition of Boson Sampling distributions, we propose using machine learning to teach computers to identify important pathologies that may be present in Boson Samplers. We illustrate the power of this approach by focusing on determining whether the bosons that were fed into the Boson Sampler are indistinguishable, as required to be, or distinguishable. Building on results of Wang and Duan [37], we train a model for identifying such pathologies. Our model consists in a compatibility test, based on data clustering, performed on experimental sampled data drawn from an untrusted Boson Sampler as well as a trusted device (see Fig. 1). We train this model using numerical simulations of small-scale Boson Sampling experiments and then evaluate its performance on larger Boson Samplers. Moreover, we experimentally test our trained model both on traditional Boson Sampling and scattershot Boson Sampling, the latter being an advanced scheme, more suitable to achieve the regime of quantum supremacy with current techonlogies. We find that, even when tested on Boson Samplers the model had never seen, the trained algorithm accurately predicts the correct result whereas prior methods did not. Furthermore, we adopted our approach to validate against other failure modes that were not considered in the training stage, thus showing the capability of machine learning techniques to detect common features in large-size datasets. Finally, we provide a detailed discussion on the physical mechanisms behind the correct functioning of a generic clustering algorithm in Boson Sampling validation. Indeed, we show that the key ingredient lies in the presence of strong correlations within the distributions obtained with indistinguishable and distinguishable photons. Such correlations correspond to a marked internal structure in the distributions, while distributions from different particle types present highly uncorrelated structures. Thanks to their inherent versatility and their capability of operating without an in-depth knowledge of the physical system under investigation, clustering techniques may prove to be effective even in a scope broader than the Boson Sampling problem [7–11].

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 [6] that corresponds to sampling from the output probability distribution obtained after the evolution of N identical, i.e. indistinguishable, bosons through a 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 [6, 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 [38]. 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 $\mathcal{A}_U(S, T) = \text{per}(U_{S,T})/(s_1! \ldots s_N! \, t_1! \ldots t_N!)^{1/2}$, where $\text{per}(U_{S,T})$ 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 [39].

In contrast with other computational problems, such as factoring [1], the validation of a quantum device solving the Boson Sampling problem is difficult because the answer cannot be tested using a classical computer. Indeed, a severe limitation on this aspect is imposed by the complexity of a complex matrix's permanent, since even the assessmnent of its correctness is a hard computational task. Thus, 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 necessary to develop suitable techniques, inherently tailored to deal with a large amount of data.

Validation with pattern recognition techniques — 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 approach in this context is provided by the field of machine learning, which studies how a computer can acquire information from input data and learn throughout the process how to make data-driven predictions or decisions [40]. Significant progresses have been achieved in this area over the past few years [41, 42]. 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. The main unsupervised learning approach is clustering, where data are grouped in different classes according to collective properties recognized by the algorithm. Several clustering methods are widely employed [43], such as *K-means* and *Hierarchical clustering*. Since these approaches designed to identify hidden patterns in a large amount of data, they 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 an arbitrary clustering method within the protocol, which allows us to choose the method to optimize performance on 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 struc-
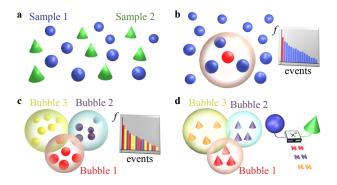
FIG. 2. **Bubble clustering validation scheme. a**, A sample is drawn from each of the two Boson Samplers to be compared. **b**, The events belonging to one of the two samples are sorted according to their observation frequency. The state with highest frequency is chosen as the center of the first cluster. Those events with distance d from the center smaller than a cut-off radius ρ_1 are included in the cluster. **c**, 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. **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. A χ^2 test with $\nu = N_{\text{bubbles}} - 1$ degrees of freedom is performed to compare the number of events belonging to the first and second sample by using the obtained cluster structure. This variable quantifies the compatibility between the samples.

ture 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 $\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{2} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$, where index j refers to the samples, index i to the clusters, n_{ij} is the number of events in the i-th cluster belonging to the j-th sample and E_{ij} is the expected value of observed events belonging to the j-th sample in the i-th cluster $E_{ij} = n_i n_j / n$, with $n_i = \sum_{j=1}^{2} n_{ij}$, $n_j = \sum_{i=1}^{N} n_{ij}$ and $n = \sum_{i=1}^{N} \sum_{j=1}^{2} n_{ij}$. 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 $\nu = N_{\text{clusters}} - 1$ degrees of freedom. 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 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 i-th mode.

Adopted clustering techniques — Several clustering methods were employed within our validation scheme: (a) a recent proposal by Wang and Duan [37], 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 were also examined, to increase the strenght of our model. A short de-

scription of each adopted method follows briefly.

(a) The protocol proposed by Wang and Duan [37], 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 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 pre-determined by the user. In the present case, the algorithm halts when no more than 1% of the observed events is included in some cluster containing less than 5 events (See Supplementary Information). 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 clustering algorithm where the user has to determine the number of classes (k) [44–46]. 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 elements coordinates. These two operations are repeated until the structure is stable. Given a set of k centroids ($c_1, ... c_k$) made of ($n_1, .. n_k$) elements ($e_{11}, ... e_{1n_1}, ..., e_{k1}, ..., e_{kn_k}$), where $\sum_{i=1}^{k} n_i = N$, the operations of selecting elements and moving the centroids minimize the objective function $\frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n_k} d(e_{ij}, c_j)$. Several trials were 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 chose to balance the two needs of clusters made up of at least 5 elements, since the compatibility test requires a χ^2 evaluation, and of a high efficacy of the validation test (See Supplementary Information).

Variations of K-means clustering — During the optimization process of *K-means clustering*, we observed that 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 considered 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 has set 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 inizialized. 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 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 to get 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.

Numerical results — As a first step, we performed a detailed analysis to identify the best algorithm among the mentioned clustering methods. More specifically, we proceeded with the two following steps: (i) a *training stage* and (ii) a *validation stage*. The parameter quantifying the capability of each test to perform correct decisions is the success percentage or, equally, 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 *training stage*, we applied all the algorithms on a *training set* of numerically generated samples of output states, belonging to the collision free subspace of $N = 3$ photons evolved through a fixed unitary transformation, with $m = 13$ modes. Hence, the dimension of the Hilbert space in this case is $\binom{13}{3} = 286$. Each algorithm was run several times, while varying the number of sampled events within the tested samples. For each considered sample size, the parameters proper of each technique were optimized. All the approaches have a common parameter, that is the minimum number n of elements (sampled output events) belonging to a cluster. Specifically, *bubble clustering* requires to determine the optimal minimum cut-off radius, *Hierarchical clustering* needs to set the maximum acceptable fraction of outliers (events belonging to clusters with less than n elements), while *K-means* requires the optimization of the number of clusters k. All algorithms have been tested by employing the L_1 and the L_2 distances. Finally, when applying majority voting, it is necessary to choose the number of distinct trials. To evaluate the success percentages for each configuration of parameters, we numerically generated 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 states, characterized by the same input mode occupation list, are evolved with the same unitary transformation U. We have performed two compatibility tests for each data set: the first between two compatible samples and the third between two incompatible ones. The results of this analysis are shown in Tab. I for sam-

Output Classification

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

TABLE I. **Confusion matrix for different clustering techniques and fixed unitary evolution (training stage).** 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 was investigated in its standard version, and initialized by *K-means++* or a preliminary run of *Hierarchical clustering*. Then, majority voting was performed on the non-deterministic versions of *K-means*. The reported success percentages were evaluated through numerical simulations by keeping the unitary evolution operator fixed. This choice is motivated by the need of training the different algorithms in order to subsequently classify new data sets.

ples of 500 output events. 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 the *K-means* approach is outperforming *bubble clustering* lies in the learning capability of *K-means*. Indeed, due to its convergence properties through the consecutive iterations, *K-means* gradually improves its insight into the internal structure that characterize the data. This feature enables a better discrimination between compatible and incompatible samples (See Supplementary Information).

(ii) In the *validation stage*, we validated the algorithm with the highest success percentage according to the results of Tab. I. We divided this task into two steps, by first (ii.a) validating its functioning for general unitary transformations and then (ii.b) by increasing the dimension of the Hilbert space. Hence, (ii.a) we performed the test with $N = 3$ photons evolving through 20 different Haar-random 13×13 matrices. For each matrix we performed 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 Tab. II, we report the means and standard deviations of the suc-

Output Classification

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

TABLE II. **Confusion matrix for *bubble clustering* and *K-means++* with majority voting random unitary evolution [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 were evaluated through numerical simulations, by drawing 20 Haar-random unitary transformation, and by adopting the same parameters obtained from stage (i) corresponding to the results of Tab. I.

cess 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 employing the L_2 distance, permits to achieve better success percentages. Then (ii.b) we tested the capability of the chosen validation method to successfully operate on Hilbert spaces with larger dimensions. More specifically, we progressively increased the number of photons and of modes, showing that the adopted test permits to validate the data samples even for larger values of N and m (see Tab. III). To extend our investigations to larger-dimensional Hilbert spaces, we exploited a recent algorithm developed by Neville *et al.* [47] to sample with a much more efficient approach compared to the brute force one. Specifically, the algorithm employs a Markov chain Monte Carlo to simulate Boson Sampling without evaluating the whole output distributions. Note that, for all dimensions of the Hilbert spaces probed by our analysis, a sample size of 6000 events is sufficient to successfully apply the protocol. For $N = 7$ and $m = 70$, this value is a small fraction (6×10^{-6}) of the number of available output combinations meaning that most of output states do not appear in the measured sample. An aspect of our test that is worth noticing, as shown in Tab. III, is that the probability of error is lopsided. This 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 the parameters proper of each technique determined in the training process for low $N = 3$ and $m = 13$, are shown to be effective for larger number of photons and circuit size. This means that the developed technique does not require to perform a different training stage for each pair (N,m).

During the validation stage, we have also performed numerical simulations to verify whether the present approach is effective against other possible failure modes different from distinguishable particles, namely the Mean-Field sampler [34] and a Uniform sampler, see Section 6 of the SI. The former performs sampling from a suitable tailored single-particle distribution which reproduces same features of multiphoton interference, while the latter performs sampling from a uniform distribution. Again, we did not repeat the training stage and thus we employed the same parameters obtained for $N = 3$, $m = 13$ with distinguishable particles. We observe that the test shows the capability to distinguish between a *bona fide* Boson Sampler and a Uniform or Mean-Field sampler. This 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 parameters, into situations where the data come from approximations to the Boson sampling distribution that prima facie bear no resemblance to the initial training examples.

Experimental results — Through the validated experimental apparatus shown in Fig. 3a, we collected data samples corresponding to the Boson Sampling distribution, and distinct samples obtained with distinguishable particles. The degree of distinguishability between the input photons is adjusted by modifying their relative arrival times through delay lines (see Methods). The unitary evolution is implemented by an integrated photonic chip realized exploiting the 3D-geometry capability of femtosecond laser writing [48] and performs the same transformation U employed for the numerical results of Tab. I. We then performed 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. 3b, for the case of incompatible samples. This implies that the reported percentages represent the capability of the test to recognize Boson Sampler fed with distinguishable photon inputs. Reshuffling of the experimental data was used to have a sufficient number of samples to evaluate the success percentages (see Supplementary Information). Hence, the tests were performed on samples drawn randomly from the experimental data.

Generalization for scattershot Boson Sampling — The scattershot version of Boson Sampling [19] is implemented through the setup of Fig. 4a. 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 (6,8) 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 considered a generalization of the proposed algorithm to be applied for scattershot 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 would bring n distinct chi-square variables χ_i^2, where the i-th variable would quantify the agreement be-

Output Classification

Modes (m)	Photons (N)						
	3		**4**		**5**		
	Ind.	Dis.	Ind.	Dis.	Ind.	Dis.	
13	20	0	20	0	20	0	Ind
	0	20	0	20	0	20	Dis.
20	20	0	20	0	20	0	Ind
	0	20	0	20	0	20	Dis.
30	20	0	20	0	19	1	Ind
	0	20	0	20	0	20	Dis.
40	20	0	19	1	19	1	Ind
	5	15	1	19	2	18	Dis.
50	20	0	19	1	20	0	Ind
	3	17	1	19	1	19	Dis.

	Ind.	Dis.	
Overall performance	296	4	Ind.
	13	287	Dis.

Output Classification

Modes (m)	Photons (N)						
	7		**6**		**5**		
	Ind.	Dis.	Ind.	Dis.	Ind.	Dis.	
50	8	2	9	1	10	0	Ind.
	0	10	0	10	0	10	Dis.
70	10	0	10	0			Ind.
	2	8	0	10			Dis.

TABLE III. **Confusion matrix for *K-means++* with majority voting by varying N and m [validation stage, step (ii.b)]** Upper table: number of successes for *K-means*, initialized by *K-means++* and with majority voting, obtained for different values of the number of photons N and modes m, the sample size here is of 6000 events. Again, we adopted the same parameters obtained from stage (i) corresponding to the results of Tab. I. Numerical samples of *bona fide* Boson Samplers were generated through brute force sampling. Middle table: overall confusion matrix obtained by summing the results over the number of photons and modes of the Upper table. Lower Table: number of successes for *K-means*, initialized by *K-means++* and with majority voting, obtained for higher-values of the number of photons N and modes m, the sample size is of 6000 events. We adopted the same parameters obtained from stage (i) corresponding to the results of Tab. I. Numerical samples of *bona fide* Boson Samplers were generated through a Markov Chain Monte Carlo independent sampler, adopting the method proposed by [47].

tween 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
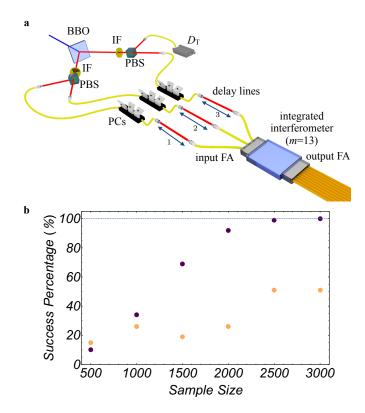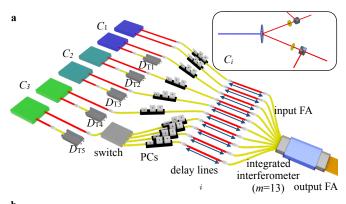

FIG. 3. **Experimental validation of a Boson Sampling experiment with N=3 and m=13. a**, Experimental setup for a $N = 3$ Boson Sampling experiment in an integrated $m = 13$ interferometer (see Methods). BBO - beta barium borate crystal; IF - interferential filter; PBS - polarizing beam-splitter; PC - polarization controller; FA - fiber array. **b**, Success percentages of the compatibility test performed on incompatible experimental data sets with different size (indistinguishable against distinguishable photon inputs), for a significance level of 5%. The darker dots represent the performance of *K-means++* with majority voting clustering algorithm, while the lighter ones were obtained with *bubble clustering*. In both cases we adopted $d = L_2$. The input state was characterized by injecting single photons in waveguides (6,7,8). The discrepancy from the numerical results of Tab. II, whose scenario is the same, are attributed to the non perfect indistinguibility of the injected photons [50].

would be tested separately. In order to extract only one parameter 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 $\nu = \sum_{i=1}^{n} \nu_i$ degrees of freedom, provided that the χ_i^2 are independent. We have performed this generalized test on the experimental data by adopting the same clustering technique previously discussed in the single input case.

Experimental results for scattershot Boson Sampling — We have collected output samples given by 8 different inputs both with indistinguishable photons and with distinguishable ones. Through the evaluation of the new variable $\tilde{\chi}^2$, the algorithm was able to distinguish between a trustworthy scattershot 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.4b. The standard version of
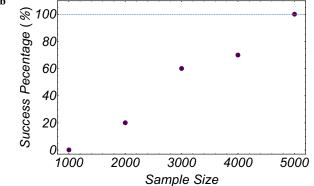
FIG. 4. **Experimental validation of a scattershot Boson Sampling experiment with N=3 and m=13**. **a**, Experimental apparatus for a $N = 3$ scattershot Boson Sampling experiment in an integrated $m = 13$ interferometer (see Methods). The input state is generated probabilistically by six independent parametric down-conversion sources (boxes in the figure) located in 3 different BBO crystals C_i (see inset). PC - polarization controller; FA - fiber array. **b**, Success percentages obtained by applying the generalized version of the compatibility test performed on incompatible experimental data samples of different size, (indistinguishable against distinguishable photon inputs) for a significance level of 5%. The adopted clustering algorithm is *K-means*, inizialized by *K-means++*, with majority voting. Experimental data sets correspond to 8 different inputs. The number of events belonging to each input state randomly varies for each sample size drawn from the complete data set.

the test, validating each input separately, would require samples of 2000 events per input to reach a success percentage $\geq 80\%$, that is, an overall amount of 16000 events. Hence, the generalized version of the test permits to significantly reduce the amount of necessary resource to validate scattershot Boson Sampling experiments.

Structure of the probability distributions — Our previous discussion has conclusively shown that, at least for the values of (N, m) considered, K–means clustering algorithms are highly effective at discriminating between samples drawn from boson samplers that use distinguishable photons versus those with indistinguishable ones. Here we provide further analysis that shows why our approach is so 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 mecha-

nism 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 (here, the L_1- and L_2-norm have been employed). The key observation we make is that the number of events necessary to effectively discriminate the input data samples is dramatically lower than one might expect. Indeed, our results have shown that, for the investigate range of values (N, m), the number of necessary events to reach a success rate near unity is almost constant (~ 6000) for increasing problem size. More specifically, the number of events for a sufficiently large number of photons N and modes m is much lower than the number of available output combinations. For instance, even for $N = 4$ and $m = 40$ the Hilbert space dimension is 91390, and thus the number of events to successfully apply the test is approximately $\sim 6.6\%$ of the full Hilbert space. By increasing the system size, such a fraction drops fast to smaller values (for instance, $\sim 0.28\%$ for $N = 5$ and $m = 50$). Accordingly the output sample from the device will mostly consist of output states occurring with no repetition. Hence, only the configurations presenting higher probability will 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 first step towards this goal, we computed 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. Fig. 5a 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 computed 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 suggest that the two distributions have different supports over the outcome distributions. The same analysis have been performed for $N = 5$ and $m = 50$, showing that a similar behavior is obtained for increasing size (see Fig. 5b). By averaging over
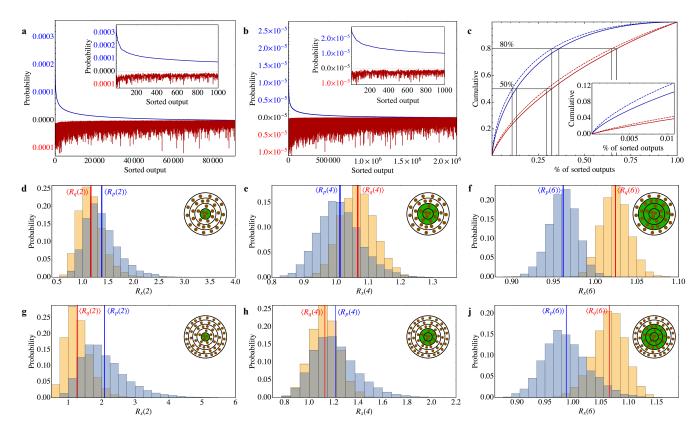
FIG. 5. **Analysis of the structure of the distributions**. **a,b**, Probability distributions for a fixed unitary U in the case of indistinguishable (blue) and distinguishable (red) photons. **a**, $N = 4$, $m = 40$, and **b**, $N = 5$, $m = 50$. The distributions are sorted by following the same ordering, so as to have decreasing probability values for the indistinguishable distribution P_i. Inset: zoom corresponding to the 1000 most probable output states. **c**, Cumulative distributions for indistinguishable (blue) and distinguishable (red) photons, by following the same ordering of panels **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: zoom corresponding to 0.01% most probable output states. **d-j**, 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 $\leq k$. **d,f**, $N = 4$, $m = 40$. **g-j**, $N = 5$, $m = 50$. Vertical lines correspond to the averages $\langle R_x(k) \rangle$, with $x = p$ (blue) and $x = q$ (red). **d,g**, $k = 2$, **e,h**, $k = 4$ and **f,j**, $k = 6$. Insets: schematic view of the spheres at distance $\leq k$, represented by concentric circles, where states are represented by brown points.

$M' = 100$ 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 $m = 40$, while being $r \sim 0.57 \pm 0.03$ 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. 5c), where the distinguishable case is sorted by following the same order of the indistinguishable one. We observe that, in order 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 $\sim 13\%$ of the overall number of outputs for indistinguishable photons, while $\sim 32\%$ are necessary for the distinguishable case (by following the above mentioned ordering procedure). Similar numbers are obtained for larger dimensionalities ($\sim 11\%$ and $\sim 30\%$ 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 that the number of combinations is collected, the outcomes actually occurring in the data sample will present lower distance values thus justifying the application of a clustering procedure.

We further probe how these correlations become visible through a clustering method by we 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 calculated the probabilities P_j and Q_j for both cases (indistinguishable and distinguishable photons) and then sorted the distribution P_j in decreasing order. Let us call J the outcome with the highest P_j value which is to say $J = \text{argmax}(P_j)$. Let us for simplicity fix the distance to be the L_1-norm (analogous results are ob-

tained for the L_2-norm). Note that the L_1-norm defined in the main text has only N possible non-trivial values $k = 2s$, with $s = 1, \ldots N$. We then estimated 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 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 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 Fig. 5d-f for $N = 4$, $m = 40$ with $k = 2, 4, 6$ (being $k = 8$ a trivial one, which includes all output states given 4-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, $\sim 0.16\%$ for $k = 2$, $\sim 4.3\%$ for $k = 4$ and $\sim 35.5\%$ for $k = 6$). Similar results have been obtained also for $N = 5$ and $m = 50$ (see Fig. 5g-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 states 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 is why basing a classifier around this effect 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 [32] 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, will 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.

To summarize, all these analyses show that Boson Sampling distributions with indistinguishable and distinguishable particles present an internal structure that can be catched by the clustering procedure at the basis of our validation method, thus rendering our method effective to discriminate between the two hypotheses.

Discussion — In this article we have shown that pattern recognition techniques can be exploited to identify pathologies in Boson Sampling experiments. The main feature of the devised approach relies in the absence of any permanent evaluation, thus not requiring the calculation of hard-to-compute quantities during the process. The efficacy of this method 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. This approach is scalable to larger Hilbert spaces and so it is a promising approach for the validating mid-term experiments. Moreover, our experimental demonstration shows that it is possible to successfully validate Boson Sampling even in lossy scenarios, which have already been shown to maintain the same computational hardness of the original problem [49].

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 had 100% classification accuracy for the training data but had roughly 95% accuracy in the test data. Had we only reported 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.

Finally, although our work is focused on 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 classifiers to learn facts about complex quantum systems that humans can easily miss. 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.

METHODS

Experimental apparatus. The laser source of the experiment generates a pulsed 785 nm field, which is frequency doubled by exploiting a second-harmonic generation (SHG) process in a BiBO (beta bismute borate) crystal, generating a 392.5 nm pump beam. In the Boson Sampling experiment [17], the pump is injected in a 2-mm thick BBO (beta barium borate) crystal cut for type-II phase matching. Four photons are produced by a second-order process, are spectrally selected by a 3-nm interference filter and are spatially separated according to their polarization state by a polarizing beam-splitter. One of the four photons is directly detected with an avalanche photodiode (detector D_T in Fig. 3) and acts as a trigger for the experiment. The other three-photons are coupled into single-mode fibers, are prepared in the same polarization state by fiber polarization controllers (PC) and propagate through independent delay lines that are employed to adjust their relative delay. Finally, they are injected in input modes (6,7,8) of a 13-mode integrated interferometers by means of a single-mode fiber array (FA), and are collected after the evolution by a multi-mode FA. The output modes are then measured with avalanche photodiodes. The output signals are elaborated by an elec-

tronic acquisition system to identify three-fold coincidences at the output of the device, conditioned to the detection of the trigger photon at D_T.

In the scattershot Boson Sampling experiment [19], the pump beam is injected into six independent down-conversion sources, physically obtained with three 2-mm long BBO crystals. In crystal C_1, the two output photons of one pair after separation by the PBS are directly injected into input modes 6 and 8 of the integrated interferometer. The third photon is obtained proba-

bilistically from the other pair generated by crystal C_1 and from the 4 pairs generated by crystals C_2 and C_3. More specifically, the third photon is injected in a variable input mode identified by the detection of the twin photon in the corresponding trigger detector D_{Ti}. Furthermore, an optical switch is employed to increase the input state variability. The sampled input states are thus of the form (6,j,8), with $j = \{1, 2, 3, 7, 9, 11, 12, 13\}$, for an overall set of 8 different input states.

---

[1] Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Scientific and Statistical Computing* **26**, 1484 (1997).

[2] Shuld, M., Sinayskiy, I. & Petruccione, F. An introduction to quantum machine learning. *Contemporary Physics* **56**, 172-185 (2015).

[3] Biamonte, J. *et al.* Quantum machine learning. *Nature* **549**, 195-202 (2017).

[4] Feynman, R. Simulating physics with computers. *Journal of Theoretical Physics* **21**, 467-488 (1982).

[5] Lloyd, S. Universal quantum simulators. *Science* **273**, 1073-1078 (1996).

[6] Aaronson, S. & Arkhipov, A. The computational complexity of linear optics. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, 333-342 (ACM, 2011).

[7] Bremner, M. J., Montanaro, A. & Shepherd, D. J. Average-case complexity versus approximate simulation of commuting quantum computations. *Phys. Rev. Lett.*, **117**, 080501 (2016).

[8] Boixo, S. *et al.* Characterizing quantum supremacy in near-term devices. ArXiv:1608.00263 (2016).

[9] Gao, X., Wang, S.-T. & Duan, L.-M. Quantum supremacy for simulating a translation-invariant Ising spin model. *Phys. Rev. Lett.* **118**, 040502 (2017).

[10] Bremner, M. J., Montanaro, A. & Shepherd, D. J. Achieving quantum supremacy with sparse and noisy commuting quantum computations. *Quantum* **1**, 8 (2017).

[11] Bermejo-Vega, J. *et al.* Architectures for quantum simulation showing a quantum speedup. ArXiv:1703.00466 (2017).

[12] Broome, M. A. *et al.* Photonic boson sampling in a tunable circuit. *Science* **339**, 794-798 (2013).

[13] Spring, J. B. *et al.* Boson sampling on a photonic chip. *Science* **339**, 798-801 (2013).

[14] Tillmann, M. *et al.* Experimental boson sampling. *Nature Photonics* **7**, 540-544 (2013).

[15] Crespi, A. *et al.* Integrated multimode interferometers with arbitrary designs for photonic boson sampling. *Nature Photonics* **7**, 545-549 (2013).

[16] Spagnolo, N. *et al.* General rules for bosonic bunching in multimode interferometers. *Phys. Rev. Lett.* **111**, 130503 (2013).

[17] Spagnolo, N. *et al.* Experimental validation of photonic boson sampling. *Nature Photonics* **8**, 615-620 (2014).

[18] Carolan, J. *et al.* On the experimental verification of quantum complexity in linear optics. *Nature Photonics* **8**, 621-626 (2014).

[19] Bentivegna, M. *et al.* Experimental scattershot boson sampling. *Science Advances* **1**, e1400255 (2015).

[20] Carolan, J. *et al.* Universal linear optics. *Science* **349**, 711-716 (2015).

[21] Loredo, J. C. *et al.* Bosonsampling with single-photon fock states from a bright solid-state source *Phys. Rev. Lett.* **118**, 130503 (2017).

[22] Wang, H. *et al.* High-efficiency multiphoton boson sampling. *Nature Photonics* **11**, 361-365 (2017).

[23] He, Y. *et al.* Time-Bin-Encoded Boson Sampling with a Single-Photon Device. *Phys. Rev. Lett.* **118**, 190501 (2017).

[24] Shen, C., Zhang, Z. & Duan, L. Scalable implementation of boson sampling with trapped ions. *Phys. Rev. Lett.* **112**, 050504 (2014).

[25] Gogolin, C., Kliesch, M., Aolita, L. & Eisert, J. Boson-sampling in the light of sample complexity (2013). ArXiv:1306.3995.

[26] Aaronson, S. & Arkhipov, A. Bosonsampling is far from uniform. *Quantum Information & computation* **14**, 1383-1423 (2014).

[27] Wiebe, N. *et al.* Using quantum computing to learn physics. *Bulletin of EATCS* **1** (2014).

[28] Hong, C. K., Ou, Z. Y., & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett.* **59**, 2044-2046 (1987).

[29] Menssen, A. J. *et al.* Distinguishability and Many-Particle Interference. *Phys. Rev. Lett.* **118**, 153603 (2017).

[30] Bentivegna, M. *et al.* Bayesian approach to boson sampling validation. *Int. J. Quantum Inform.* **12**, 1560028 (2014).

[31] Shchesnovich, V. Universality of generalized bunching and efficient assessment of boson sampling. *Phys. Rev. Lett.* **116**, 123601 (2016).

[32] Walschaers, M. *et al.* Statistical benchmark for bosonsampling. *New J. Phys.* **18**, 032001 (2016).

[33] Bentivegna, M., Spagnolo, N. & Sciarrino, F. Is my boson sampler working? *New J. Phys.* **18**, 041001 (2016).

[34] Tichy, M. C., Mayer, K., Buchleitner, A. & Molmer, K. Stringent and efficient assessment of boson-sampling devices. *Phys. Rev. Lett.* **113**, 020502 (2014).

[35] Crespi, A. Suppression law for multiparticle interference in sylvester interferometers. *Phys. Rev. A* **91**, 013811 (2015).

[36] Crespi, A. *et al.* Suppression law of quantum states in a 3d photonic fast fourier transform chip. *Nature Communications* **7**, 10469 (2016).

[37] Wang, S. T. & Duan, L.-M. Certification of boson sampling devices with coarse-grained measurements. ArXiv:1601.02627.

[38] Valiant, L. G. The complexity of computing the permanent. *Theor. Comput. Sci.* **8**, 189-201 (1979).

[39] Scheel, S. Permanent in linear optical networks (2004). ArXiv:quant-ph/0406127.

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

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

[42] Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT press, 2012).

[43] Rokach, L. & Maimon, O. *Data Mining and Knowledge Discovery Handbook*, chap. Clustering Methods (Springer US, 2005).

[44] MacQueen, J. Some methods for classification and analysis of multivariate observations. In of Calif. Press, U. (ed.) *Proc.*

Fifth Berkeley Symp. on Math. Statist. and Prob., vol. 1, 281-297 (1967).

[45] Biometrics, E. F. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**, 768-769 (1965).

[46] Lloyd, S. P. Least squares quantization in pcm. *IEEE Transactions on Information Theory* **28**, 129-137 (1982).

[47] Neville, A. *et al.* Classical boson sampling algorithms with superior performance to near-term experiments *Nature Physics*, advance online publication, doi:10.1038/nphys4270 (2017).

[48] Gattass, R. & Mazur, E. Femtosecond laser micromachining in transparent materials. *Nature Photonics* **2**, 219-225 (2008).

[49] Aaronson, S. *et al.* BosonSampling with lost photons. *Phys. Rev. A* **93**, 012335 (2016).

[50] Spagnolo, N. *et al.* Three-photon bosonic coalescence in an integrated tritter. *Nature Communications* **4** 1606 (2013).

ACKNOWLEDGMENTS

Supplementary Information
Pattern recognition techiques 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, WA, United States*

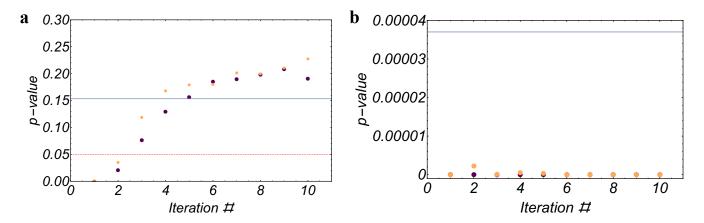## SUPPLEMENTARY NOTE 1: EFFICIENCY OF THE DIFFERENT PATTERN RECOGNITION TECHINQUES

Let us analyze in more detail the characteristics of two of the different pattern recognition techniques that we applied to Boson Sampling validation, i.e. *K-means* and *bubble clustering*. Specifically, the main strength of *K-means clustering* [S1] lies in the ability of the algorithm to learn, through a certain number of iterations, a locally optimal cluster structure for the experimental data to be analyzed. This feature is clearly highlighted when we compare the performances of the compatibility test, that is the resulting p-values, when we adopt *K-means clustering* and *bubble clustering*. In Supplementary Fig. 1 a we compare two incompatible samples, i.e. one drawn from a trustworthy Boson Sampler and one from a Boson Sampler fed with a distinguishable three photon input. In Supplementary Fig. 1 b we instead compare two compatible samples, both drawn from a trustworthy Boson Sampler. *Bubble clustering* is not iterative and the resulting cluster structure is deterministic for a given data set. Indeed, the algorithm chooses the cluster centers accordingly to the observation frequency of the output states and assigns the elements only once. On the other hand, *K-means* convergence is not univocal, since the final cluster structure depends on the centroids inizialization, and, through the selection of the centroids and the corresponding assigned elements, it is guaranteed to locally optimize its objective function, $\frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n_k} d(e_{ij}, c_j)$ [S2]. Having considered this feature of *K-means*, we performed 200 tests both on the two compatible samples and on the incompatible ones, using 100 different uniformly drawn random inizializations, and 100 *K-means++* inizializations. The pattern recognition techniques based on *K-means*, given its objective function convergence to local minima property, improve the test's result after each iteration, being able to recognize whether the samples are compatible or not with better results than *bubble clustering*, that is, leading to a smaller p-value, with incompatible samples and a greater p-value with compatible samples. The p-value quantifies the significance of the χ^2 test result, since it gives the probability of obtaining a greater value of the χ^2 variable, if the null hypothesis of compatible samples is true.

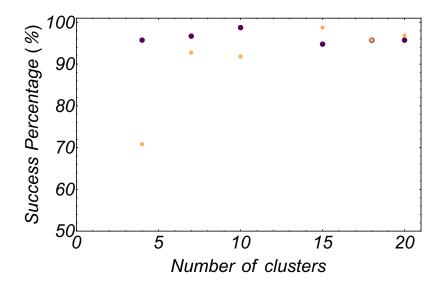## SUPPLEMENTARY NOTE 2: FINDING THE OPTIMAL NUMBER OF CLUSTERS FOR K-MEANS

All clustering techniques investigated in the present study are characterized by several parameters that have to be tuned to ensure the right operation for the algorithm. *K-means clustering*, in particular, requires the user to set the number of clusters forming the structure. The approach we used to set this parameter in the training stage consisted in multiple trials of the compatibility test on sets of the same size and varying the number of clusters. The success percentage of the test increases significantly as the number of cluster is raised, especially in the case of incompatible samples, as shown in Supplementary Fig. 2. The increase in the number of clusters was halted by requiring that clusters are composed of at least 5 elements, to ensure the correct operation of the compatibility test (that requires the evaluation of a χ^2 variable).

SUPPLEMENTARY NOTE 3: HALTING CONDITION FOR HIERARCHICAL CLUSTERING

The choice of the halting condition for *Hierarchical clustering* attempts to balance two conflicting requirements: the number of clusters and their minimum size. Firstly, there is the need to have clusters with at least 5 elements, to make the χ^2 test meaningful. However, the minimum size is not the optimal choice as halting condition, since there is no control over the number of clusters and there is the risk that the algorithm stops with too few clusters remaining. An excessively low number of clusters, less than 3, can compromise the correct operation of the test. We therefore chose to remove outlier elements in the data, neglecting them in the final cluster structure. This is done by requiring that the algorithm stops when a predetermined fraction of the elements belongs to clusters with less than 5 elements. The fraction was then tuned to maximize the efficiency of the validation test, while monitoring the amount of data taken away.

Supplementary Figure 1. **K-means clustering vs bubble clustering**. **a)**, P-values obtained by the application of our compatibility test on two compatible samples of 2000 events, both drawn from a trustworthy Boson sampler, with N=3 and m=13. The blue horizontal line indicates the result obtained by using the cluster structure given by *bubble clustering*. The darker and lighter dots represent the p-values obtained respectively from the application of the test on the cluster structures obtained by *K-means*, initialized by uniformly drawn random centroids and by *K-means++*, at each iteration. Since the obtained cluster structure is not deterministic, we performed the mean on the p-values corresponding to 100 different inizializations. The red dashed line indicates the p-value above which the two samples are recognized as compatible. **b)**, P-values obtained by the application of our compatibility test on two incompatible samples of 2000 events, one drawn from a trustworthy Boson sampler and the other from a Boson Sampler fed with a distinguishable photon input. The blue horizontal line indicates the result obtained via *bubble clustering*. The darker and lighter dots represent, respectively, the p-values obtained by *K-means*, initialized by uniformly drawn random centroids and via *K-means++* inizialization. Since the cluster structure obtained by *K-means* is not deterministic, we performed the mean on the p-values corresponding to 100 different inizializations.
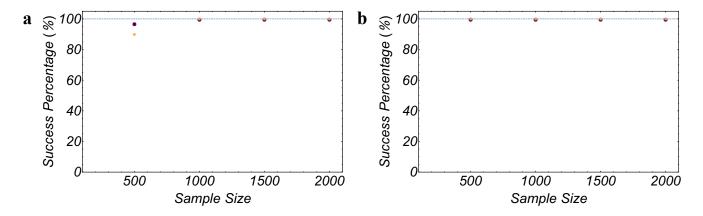

Supplementary Figure 2. **Success percentages vs number of clusters (k)**. Success Percentages of the compatibility test, with *K-means clustering*, obtained by comparing numerically generated samples of 1000 events with N=3 and m=13 drawn from the Boson Sampling distribution and from the distribution given by distinguishable photon inputs, as a function of the number of cluster. The darker and lighter dots represent the percentages corresponding respectively to compatible and incompatible samples. The percentages here indicated are evaluated considering 5% as significance level for the χ^2 test.

SUPPLEMENTARY NOTE 4: RESHUFFLING OF THE EXPERIMENTAL DATA

Our evaluation of the experimental success percentages of the validation test uses 100 sets of three data samples, two compatible and one incompatible. Having such a number of data allows to perform 100 compatibility tests between compatible samples and 100 between incompatible samples. The required sample size to ensure more than 0.8 probability to recognize two incompatible samples is 500 events. This sample size implies that, for each input mode occupation list, we would need at least 5×10^4 events drawn from the distribution with distinguishable photon input and 10×10^5 events drawn from the corresponding

Boson sampling distribution. Since the amount of experimental data available was not sufficient, we adopted a reshuffling of the experimental data. To this purpose, starting from a set of N_{exp} output events sampled experimentally, we picked randomly the number of events corresponding to the required sample size N_e among the available data, as long as $N_e < N_{exp}$. This approach may bring a small bias in the results. To quantify this bias, we performed a numerical simulation, comparing the trends of the success percentage as a function of the sample size, with and without reshuffling. The trend is the same, but the growth is slower. Specifically, as reported in Supplementary Figure 3, numerical simulations show that, for the investigated sample sizes, reshuffling of data does not affect the result of the validation algorithm adopting *K-means clustering*, initialized with *K-means++*. This analysis implies that the success percentages obtained experimentally and showed in Fig. 3 b and Fig. 4 of the main text are reliable.
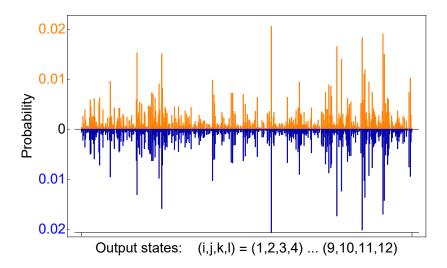

Supplementary Figure 3. **Reshuffling of the data**. **a** Success Percentages of the compatibility test performed on incompatible samples, with N=3 and m=13, with *K-means clustering* initialized with *K-means++*. The lighter and darker dots represent the percentages corresponding respectively to reshuffled and to not reshuffled data. **b** Success Percentages of the compatibility test on incompatible samples, with N=3 and m=13, adopting *K-means clustering* initialized with *K-means++*, with majority voting. The lighter and darker dots represent the percentages corresponding respectively to reshuffled and to not reshuffled data. The percentages here indicated are evaluated considering 5% as significance level for the χ^2 test.

SUPPLEMENTARY NOTE 5: PATTERN RECOGNITION TECHNIQUES ON HIGH-DIMENSIONAL HILBERT SPACES

Tab. III of the main text shows the performance of our validation protocol K-*means++ with majority voting* for a wide range of Hilbert-space dimensions (up to 10^6 in the case N=5, m=50). Tests were performed adopting the same parameters, i.e. sample size, number of clusters and number of trials for majority voting, obtained training our algorithm in a Hilbert space corresponding to N=3, m=13. The limit on the growth of the Hilbert space was imposed by the computational resources required for the numerical simulation of a *bona fide* Boson Sampler with $N > 5$ and $m > 50$. To perform the simulation of a higher-dimensional *bona fide* Boson Sampler we adopted a method recently introduced in [S4], which exploits Markov chain Monte Carlo independent sampler to approximate genuine Boson Sampling. The agreement between this approximate sampling and the genuine one (i.e. from the -known- distribution with indistinguishable particles) has been assessed using the total variation distance (see Supplementary Fig. 4) , confirming the quality of the approximation in the range of (N, m) that was possible to probe with a brute force approach. As we show Tab. III of the main text, through this algorithm we were able to simulate *bona fide* Boson Samplers up to the case of $N = 7$, $m = 70$. We then performed validation tests adopting the K *means ++ with majority voting* protocol to compare *bona fide* Boson samplers with compatible samplers and Boson Samplers fed by distinguishable-particle inputs, still using all the previously fixed parameters. As shown in Tab. III of the main text, the parameters learned for our protocol are effective for a wide range of Hilbert space dimensions, indeed tests were performed from a dimension of 10^2 to 10^9, providing evidence that in the probed range size of the samples required to perform the validation tests does not exhibit an exponential growth.

SUPPLEMENTARY NOTE 6: BOSON SAMPLING VS UNIFORM AND MEAN-FIELD SAMPLING WITH K-MEANS++

We analyzed the efficiency of our pattern recognition validation protocol also to discriminate *bona fide* Boson Samplers from uniform samplers and Mean-Field samplers. Mean-Field sampler (Mf), firstly described in [S3], is a physically plausible and

Supplementary Figure 4. **Generating samples with a Markov-chain Monte Carlo**. By adopting the Markov chain Monte Carlo method introduced in [S4], it is possible to sample with a more efficient approach than the brute force one while maintaining a good agreement with the ideal distribution with indistinguishable particles. Here, an example of distribution p^{MC} retrieved from the Markov chain for N=4 and m=12 (orange) compared to the ideal one p (blue): the Total Variation Distance, defined as TVD = $1/2 \sum |p_i^{MC} - p_i|$, is lower than 2% after a number of samples 10^2 times greater than the size of the Hilbert space.

efficiently simulable system that reproduces many of the interference features characteristic of a genuine Boson Sampler. In this scenario, with N bosons and m modes in the probability distribution is given by the following equation:

$$P_{Mf}(T_{Mf}, S_{Mf}, U) = \frac{N!}{N^N \sum_{l=1}^{N}} \sum_{q=1}^{N} \prod_{k=1}^{N} |U_{k_q j_k}|^2 \tag{S1}$$

where T_{Mf}=$(t_1, ..., t_m)$ is the output configuration and S_{Mf}=$(s_1, ..., s_m)$ the input one, with respectively $(k_1,..., k_N)$ and $(j_1, ..., j_N)$ as mode arrangements. So, for each of Hilbert space dimensions shown in Tab. III of the main text, we drew 10 samples of 6000 events from the three mentioned distributions (Boson Sampling, Mean-Field and uniform distribution), for a fixed unitary transformation U. The output sample from the Boson Sampling distribution was generated with a direct brute force approach for the lowest-dimension Hilbert spaces and employing the more efficient Markov chain Monte Carlo independent sampler for a larger number of photons and modes [S4] (see Supplementary Note 5) . We then compared samples drawn from the Boson Sampling distribution to those drawn from the alternative samplers, adopting the same parameters used in Tab. III of the main text (i.e. sample size of 6000 events and 25 clusters). As shown in Supplementary Tab. S1, the test was able to recognize the three kinds of samples as incompatible, with a considered significance level of 1%. Indeed, given the high performance of our algorithm in these tests it should come as no surprise that including majority voting on top of the previous results did not yield any further advantage. Note that the test is successfully applied to distinguish between different failure modes that were not considered in the training stage. The fact that our algorithm has never seen such proxies for the Boson sampling distribution and yet performs exceedingly well in classifying samples drawn underscores the fact that our approach yields much broader tools for validating Boson samplers than the limited scope of the training data used may suggest.

SUPPLEMENTARY REFERENCES

[S1] MacQueen, J. Some methods for classification and analysis of multivariate observations. In of Calif. Press, U. (ed.) Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., vol. 1, 281 - 297 (1967).

[S2] Bottou, L. & Bengio, Y. Convergence properties of the k-means algorithms. In Advances in Neural Information Processing Systems, 7 (MIT press, 1995).

[S3] Tichy, M. C. & Mayer, K. & Buchleitner, A. & Molmer, K., Stringent and Efficient Assessment of Boson-Sampling Devices, Phys. Rev. Lett., vol 113, 020502 (2014).

[S4] Neville, A. & Sparrow, C. & Clifford, R. & Johnston E. & Birchall, P. M. & Montanaro, A. & Laing, A.,Classical boson sampling algorithms with superior performance to near-term experiments, Nat. Phys., advance online publication, http://dx.doi.org/10.1038/nphys4270 (2017).

	Output Classification				**Output Classification**		
(N, m)	*Ind.*	*M. F.*		**(N, m)**	*Ind.*	*Unif.*	
(3,13)	10	0	*Ind.*	**(3,13)**	10	0	*Ind.*
	0	10	*M. F.*		0	10	*Unif.*
(5,50)	10	0	*Ind.*	**(5,50)**	10	0	*Ind.*
	0	10	*M. F.*		0	10	*Unif.*
(6,50)	9	1	*Ind.*	**(6,50)**	10	0	*Ind.*
	0	10	*M. F.*		0	10	*Unif.*

Supplementary Table S1. **Boson Sampling vs Mean-Field Sampling and Uniform Sampling**. Left Table: Confusion matrix showing the results obtained by applying single trial *K-means clustering* initialized with *K-means++* to distinguish between a *bona fide* Boson Sampler and a Mean-Field Sampler. Right Table: Confusion matrix showing the results obtained by applying single trial *K-means clustering* initialized with *K-means++* to distinguish between a *bona fide* Boson Sampler and a Uniform Sampler. In both cases the parameters that were used are the same that were tuned for the case $N = 3$ and $m = 13$, which gave the results shown in Tab.III of the Main text.