doi:10.1038/nature24047

Solving a Higgs optimization problem with quantum annealing for machine learning

Alex Mott[1]†*, Joshua Job[2,3]*, Jean-Roch Vlimant[1], Daniel Lidar[3,4] & Maria Spiropulu[1]

The discovery of Higgs-boson decays in a background of standard-model processes was assisted by machine learning methods[1,2]. The classifiers used to separate signals such as these from background are trained using highly unerring but not completely perfect simulations of the physical processes involved, often resulting in incorrect labelling of background processes or signals (label noise) and systematic errors. Here we use quantum[3–6] and classical[7,8] annealing (probabilistic techniques for approximating the global maximum or minimum of a given function) to solve a Higgs-signal-versus-background machine learning optimization problem, mapped to a problem of finding the ground state of a corresponding Ising spin model. We build a set of weak classifiers based on the kinematic observables of the Higgs decay photons, which we then use to construct a strong classifier. This strong classifier is highly resilient against overtraining and against errors in the correlations of the physical observables in the training data. We show that the resulting quantum and classical annealing-based classifier systems perform comparably to the state-of-the-art machine learning methods that are currently used in particle physics[9,10]. However, in contrast to these methods, the annealing-based classifiers are simple functions of directly interpretable experimental parameters with clear physical meaning. The annealer-trained classifiers use the excited states in the vicinity of the ground state and demonstrate some advantage over traditional machine learning methods for small training datasets. Given the relative simplicity of the algorithm and its robustness to error, this technique may find application in other areas of experimental particle physics, such as real-time decision making in event-selection problems and classification in neutrino physics.

The discovery of the Higgs boson at the Large Hadron Collider (LHC)[1,2] marks the beginning of a new era in particle physics. Experimental particle physicists at the LHC are measuring the properties of the new boson[11,12], searching for heavier Higgs bosons[13] and trying to understand whether the Higgs boson interacts with dark matter[14]. Cosmologists are trying to understand the symmetry-breaking Higgs phase transition that took place early in the history of the Universe and whether that event explains the excess of matter compared to antimatter[15]. The measured mass of the Higgs boson[13] implies that the symmetry-breaking quantum vacuum is metastable[16] unless new physics intervenes. The implications of the discovery of the Higgs boson will keep motivating physics research for years to come.

One of the key requirements for precisely measuring the properties of the Higgs boson is selecting large, high-purity samples that contain the production and decay of a Higgs particle. Machine learning techniques[17] could potentially be used as powerful tools for selecting such samples, but challenges remain. These challenges are greater when an investigation requires faithful simulation not only of the physics observables themselves, but also of their correlations in the data. In the measurement of the properties of the Higgs boson[11], disagreements between simulations and observations result in label noise and systematic uncertainties in the efficiency of the classifiers that adversely effect the classification performance and translate into uncertainties on the measured properties of the discovered particle.

To address these challenges in the Higgs-signal-versus-background optimization problem, we study a binary classifier that is trained with classical simulated annealing[7,8] and quantum annealing[3–6,18]. To implement quantum annealing we use a programmable quantum annealer (D-Wave Systems, Inc.) housed at the University of Southern California's Information Sciences Institute, which comprises 1,098 superconducting flux qubits. The optimization problem is mapped to one of finding the ground state of a corresponding Ising spin model. We use the excited states in the vicinity of the ground state in the training method to improve the accuracy of the classifiers beyond the baseline ground-state-finding model. We refer to this approach as quantum annealing for machine learning (QAML).

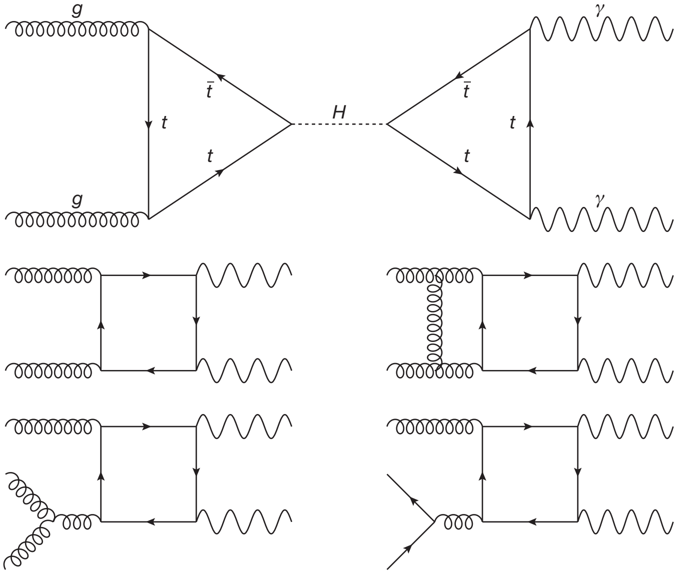

Figure 1 | Representative Feynman diagrams of processes that contribute to the simulated distributions of the Higgs signal and of the background standard-model processes. The signal corresponds to the production of a Higgs boson (H) through the fusion of two gluons (g), which then decays into two photons (γ) (top). The gluon fusion and Higgs decay processes both proceed through virtual top quark (t) loops; $\bar{t}$ is an antitop quark. Representative leading-order and next-to-leading-order background processes are standard-model two-photon production processes (bottom).

[1]Department of Physics, California Institute of Technology, Pasadena, California 91125, USA. [2]Department of Physics, University of Southern California, Los Angeles, California 90089, USA. [3]Center for Quantum Information Science and Technology, University of Southern California, Los Angeles, California 90089, USA.[4]Departments of Electrical Engineering, Chemistry and Physics, University of Southern California, Los Angeles, California 90089, USA. †Present address: DeepMind, London, UK.
*These authors contributed equally to this work.

Table 1 | The kinematic variables used to construct weak classifiers

Variable	Description
$p_T^1/m_{\gamma\gamma}$	Transverse momentum (p_T) of the photon with the larger p_T (photon '1'), divided by the invariant mass of the diphoton pair ($m_{\gamma\gamma}$)
$p_T^2/m_{\gamma\gamma}$	Transverse momentum (p_T) of the photon with the smaller p_T (photon '2'), divided by the invariant mass of the diphoton pair ($m_{\gamma\gamma}$)
$(p_T^1+p_T^2)/m_{\gamma\gamma}$	Sum of the transverse momenta of the two photons, divided by their invariant mass
$(p_T^1-p_T^2)/m_{\gamma\gamma}$	Difference of the transverse momenta of the two photons, divided by their invariant mass
$p_T^{\gamma\gamma}/m_{\gamma\gamma}$	Transverse momentum of the diphoton system, divided by its invariant mass
$\Delta\eta$	Difference between the pseudorapidity $\eta=-\log[\tan(\theta/2)]$ of the two photons, where θ is the angle with the beam axis
ΔR	Sum in quadrature of the separation in pseudorapidity η and azimuthal angle ϕ of the two photons ($\sqrt{\Delta\eta^2+\Delta\phi^2}$)
$\lvert\eta^{\gamma\gamma}\rvert$	Pseudorapidity of the diphoton system

Our criterion for comparing various classifier construction methods is the accuracy of the classifier. A classifier that is slow to train may be practically more useful than one that is less accurate but faster to train.

We model the Higgs diphoton decay channel $H\rightarrow\gamma\gamma$; see Fig. 1 for Feynman diagrams of the Higgs production and decay processes. We represent this system via the momentum of the Higgs particle, the momenta of the two photons, the angle with the beam axis θ and the azimuthal angle ϕ. More specifically, we select eight of the kinematic variables that describe the events that are generated as the variables for our classifier (see Table 1). The first five are related to the highest (p_T^1) and second-highest (p_T^2) transverse momentum (the momentum perpendicular to the axis defined by the colliding protons) of the photon pair: $p_T^1/m_{\gamma\gamma}$, $p_T^2/m_{\gamma\gamma}$, $(p_T^1 \pm p_T^2)/m_{\gamma\gamma}$ and $p_T^{\gamma\gamma}/m_{\gamma\gamma}$, where $m_{\gamma\gamma}$ is the invariant mass of the diphoton pair and $p_T^{\gamma\gamma}$ is the transverse momentum of the diphoton system. The last three are: $\Delta\eta$, the separation in the pseudorapidity $\eta=-\log[\tan(\theta/2)]$ of the two photons (η is a pseudo-invariant proxy to θ that is commonly used in high-energy physics); $\Delta R=\sqrt{\Delta\eta^2+\Delta\phi^2}$, the sum in quadrature of the separation in η and in ϕ of the two photons; and $|\eta^{\gamma\gamma}|$, the pseudorapidity of the diphoton system. In Fig. 2 we show the distribution of these variables for the signal and background datasets. The differences between these distributions are used by the classifier to distinguish the signal from the background. In addition to these eight variables, we incorporate various products between them (using rules explained in Supplementary Information) for a total of 36 (see Table 2).

We construct weak classifiers from our distributions of kinematic variables, as shown in Fig. 2 and described in Methods. We build the corresponding Ising problem as follows[6]. Let $\mathcal{I}=\{\boldsymbol{x}_\tau, y_\tau\}$ denote a set of training events labelled by the index τ, where $\boldsymbol{x}_\tau$ is a vector of the values of each of the variables that we use, and $y_\tau=\pm1$ is a binary label for whether $\boldsymbol{x}_\tau$ corresponds to signal (+1) or background (−1).

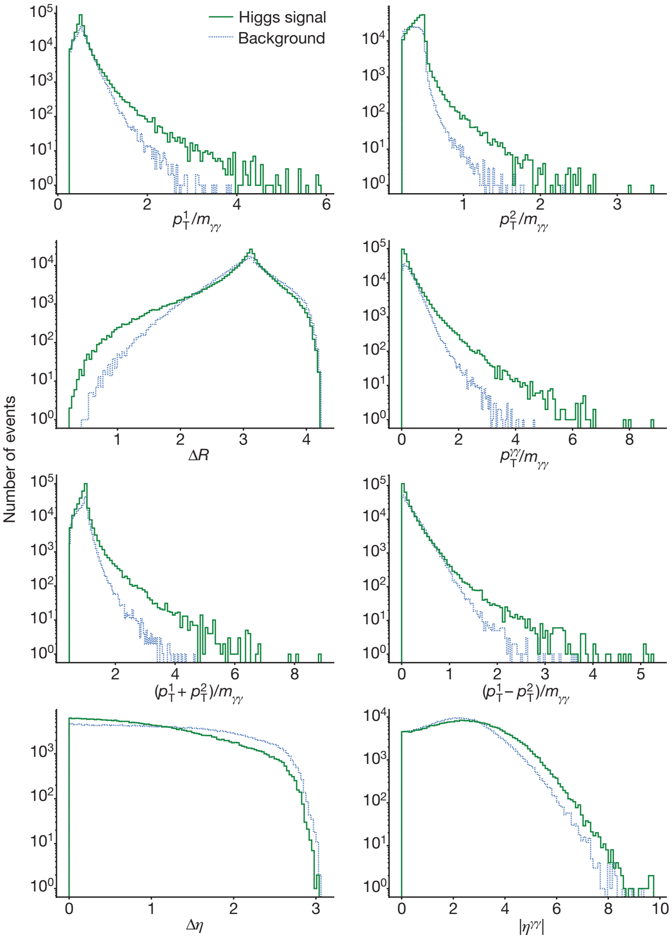

Figure 2 | Distributions of the eight kinematic variables used to construct weak classifiers. The solid green line is the signal distribution and the dotted blue line is the background. For each variable the vertical axis shows the raw count of the number of events. The total number of events simulated in each case is 307,732.

If $c_i(\boldsymbol{x}_\tau)=\pm1/N$ denotes the value of weak classifier i on the event, where N is the number of weak classifiers, equal to the number of spins or qubits, then with

$$C_{ij}=\sum_{\tau} c_i(\boldsymbol{x}_\tau)c_j(\boldsymbol{x}_\tau), \quad C_i=\sum_{\tau} c_i(\boldsymbol{x}_\tau)y_\tau$$

and a penalty $\lambda>0$ to prevent overtraining, the Ising Hamiltonian is

$$H=\sum_{i,j} J_{ij}s_i s_j+\sum_i h_i s_i$$

Table 2 | Map from number to variable or weak classifier

Number	Variable	Number	Variable	Number	Variable	Number	Variable
1	p_T^1	10	$p_T^2/(p_T^1-p_T^2)$	19	$p_T^1 p_T^2$	28	$p_T^2/p_T^{\gamma\gamma}$
2	p_T^2	11	$p_T^2/\Delta\eta$	20	$p_T^1/\Delta R$	29	$p_T^2(p_T^1+p_T^2)$
3	ΔR	12	$p_T^2\eta^{\gamma\gamma}$	21	$p_T^1/p_T^{\gamma\gamma}$	30	$(p_T^1+p_T^2)/p_T^{\gamma\gamma}$
4	$p_T^{\gamma\gamma}$	13	$1/(\Delta R p_T^{\gamma\gamma})$	22	$p_T^1(p_T^1+p_T^2)$	31	$\eta^{\gamma\gamma}/p_T^{\gamma\gamma}$
5	$p_T^1+p_T^2$	14	$(p_T^1+p_T^2)/\Delta R$	23	$p_T^1/(p_T^1-p_T^2)$	32	$1/(p_T^{\gamma\gamma}\Delta\eta)$
6	$p_T^1-p_T^2$	15	$1/\left[\Delta R(p_T^1-p_T^2)\right]$	24	$p_T^1/\Delta\eta$	33	$1/\left[p_T^{\gamma\gamma}(p_T^1-p_T^2)\right]$
7	$\Delta\eta$	16	$1/(\Delta R\Delta\eta)$	25	$p_T^1/\eta^{\gamma\gamma}$	34	$(p_T^1+p_T^2)/(p_T^1-p_T^2)$
8	$\eta^{\gamma\gamma}$	17	$\eta^{\gamma\gamma}/\Delta R$	26	$p_T^2/\Delta R$	35	$(p_T^1+p_T^2)/\Delta\eta$
9	$(p_T^1+p_T^2)\eta^{\gamma\gamma}$	18	$1/\left[(p_T^1-p_T^2)\Delta\eta\right]$	27	$\eta^{\gamma\gamma}/(p_T^1-p_T^2)$	36	$\eta^{\gamma\gamma}/\Delta\eta$

Table 3 | Variable inclusion in the ground states of instances of the Ising problem

λ	0	0.01	0.02	0.05	0.1	0.2	0.4	0.8	λ	0	0.01	0.02	0.05	0.1	0.2	0.4	0.8
1	0	0	0	0	0	0	0	0	19	20	20	20	20	20	18	0	0
2	20	20	20	20	20	20	20	20	20	0	0	0	0	0	0	0	0
3	20	20	20	20	20	20	0	0	21	0	0	0	0	0	0	0	0
4	20	20	20	20	20	20	2	0	22	19	19	19	19	1	0	0	0
5	19	19	19	19	19	19	19	0	23	0	0	0	0	0	0	0	0
6	20	20	20	20	20	20	20	0	24	20	20	20	20	20	20	7	0
7	20	20	20	20	20	20	20	9	25	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	26	3	2	1	0	0	0	0	0
9	5	4	4	1	0	0	0	0	27	0	0	0	0	0	0	0	0
10	20	20	20	20	20	20	20	18	28	20	20	20	20	20	20	20	20
11	20	20	20	20	20	14	17	0	29	19	19	19	16	1	0	0	0
12	20	20	20	20	20	20	20	0	30	7	6	4	1	0	0	0	0
13	20	20	20	20	20	20	20	20	31	0	0	0	0	0	0	0	0
14	19	19	19	19	19	12	0	0	32	15	15	15	11	5	0	0	0
15	20	20	20	20	20	20	20	2	33	0	0	0	0	0	0	0	0
16	17	17	16	10	6	4	1	0	34	19	19	19	19	16	0	0	0
17	20	20	20	20	14	1	0	0	35	20	20	20	20	20	20	20	19
18	20	20	20	17	2	0	0	0	36	20	20	20	20	20	20	3	0

The variables are listed by number (see Table 2). We show how many out of 20 training sets had the given variable turned on in the ground-state configuration. Of the 36 variables, 3 were included for all values of the penalty term λ and for all of the training sets, p_T^2, $1/(\Delta R p_T^{\gamma\gamma})$ and $p_T^2/p_T^{\gamma\gamma}$; the variables $p_T^2/(p_T^1 - p_T^2)$ and $(p_T^1 + p_T^2)/\Delta\eta$ were present in almost all; and 7 were never included, among which are the original kinematic variables p_T^1 and $\eta^{\gamma\gamma}$.

where $s_i = \pm 1$ is the ith Ising spin variable, $J_{ij} = C_{ij}/4$ is the coupling between spins i and j, and $h_i = \lambda - C_i + \frac{1}{2}\sum_j C_{ij}$ is the local field on spin i. The problem that quantum or simulated annealing attempt to solve is minimizing H and returning the minimizing, ground-state spin configuration $\{s_i^g\}_i$. The strong classifier is then constructed as

$$R(\boldsymbol{x}) = \sum_i s_i^g c_i(\boldsymbol{x}) \in [-1, 1]$$

for each new event $\boldsymbol{x}$ that we wish to classify[6]. We introduce an additional layer into our study by also constructing strong classifiers from excited-state spin configurations.

As benchmarks for traditional machine learning methods, we train a deep neural network (DNN) using Keras[9] with the Theano backend[19], and an ensemble of boosted decision trees using XGBoost (XGB)[10], using optimized choices for training hyperparameters (details of which can be found in Supplementary Information).

We compare the ground-state configurations for $\lambda \in \{0.01, 0.05, 0.1, 0.2, 0.4, 0.8\}$. A larger λ implies an increased penalty against including additional variables, and so we expect the variables included at $\lambda = 0.8$ to be determining the performance of the classifiers. Table 3 presents the relative strength of the variables in determining the performance of the classifier by showing how often variables are included in the ground-state configuration of the full 36-variable problem derived from 20 different training sets with 20,000 training events each, as a function of the penalty term λ. We find that two of the original kinematic variables, p_T^1 and $|\eta^{\gamma\gamma}|$, are never included. The number of classifiers included in the ground state of the corresponding Hamiltonian of all 20 training samples is 16 out of 36 for $\lambda \leq 0.05$ and the following three for $\lambda = 0.8$: (i) $p_T^2/m_{\gamma\gamma}$, (ii) $(\Delta R p_T^{\gamma\gamma})^{-1}$ and (iii) $p_T^2/p_T^{\gamma\gamma}$. These three classifiers have the greatest effect on the performance of the network, but would have been difficult to guess *a priori* in their composite form. The physical reason for why these variables are important for the classifier can be gleaned by considering the kinematics of the system. The key difference between an event in which a Higgs boson decays to two photons and another process that produces two photons in its final state is the production of the heavy particle in the event. A heavy particle will require considerably more energy to boost perpendicular to the beamline and hence we would expect real Higgs events to have a characteristically lower $p_T^{\gamma\gamma}$ than do background events. Because the system with the Higgs boson has less transverse boost, we would expect the two photons to have similar p_T spectra. Consequently, the second most energetic photon will typically be higher than in events without the heavy process. The p_T of the first photon is largely determined by the overall energy that is available in the collision, which is also set by $m_{\gamma\gamma}$; hence $p_T^1/m_{\gamma\gamma}$ is largely stochastic and provides little discrimination.

We estimate the receiver operating characteristic (ROC) curves on the training set and construct a final output classifier such that for

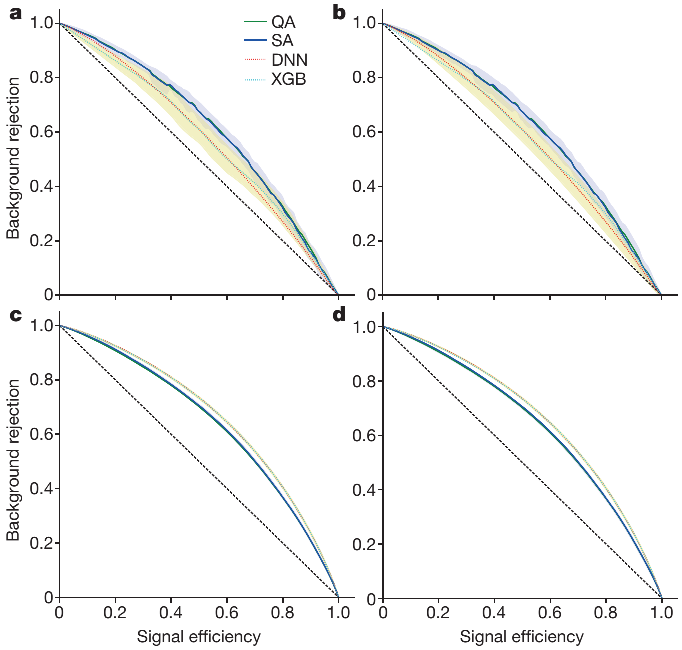

Figure 3 | Receiver operating characteristic (ROC) curves for the annealer-trained networks with $f = 0.05$, the DNN and XGB. **a–d**, Results shown are for the 36-variable networks at $\lambda = 0.05$, trained on 100 (**a** and **b**) or 20,000 (**c** and **d**) events. The ROC curve illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied, and is created by plotting the background rejection against the signal efficiency at various threshold settings. The short-dashed black line indicates no discrimination. Solid lines correspond to quantum (QA; green) or simulated (SA; blue) annealing, and dotted lines to the DNN (red) or XGB (cyan). Error bars are defined by the variation over the training sets and statistical error; 1σ error bars for quantum annealing and the DNN are shown as light blue and pale yellow shading, respectively, in **a** and **c**. The 1σ error bars for simulated annealing and XGB are included in **b** and **d**, but are too small to be visible owing to the larger number of events. For 100 events the annealer-trained networks have a larger AUROC, as shown directly in Fig. 4. The situation is reversed for 20,000 training events.

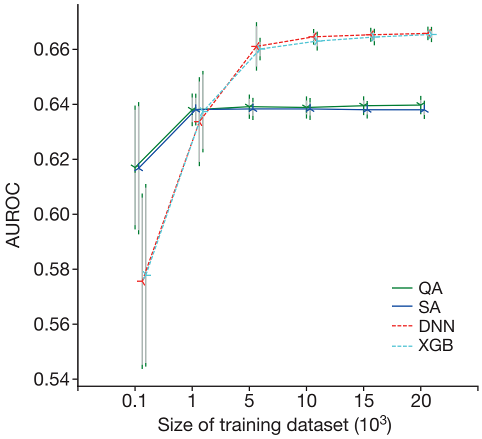

Figure 4 | Area under the ROC curve (AUROC) for the annealer-trained networks with $f = 0.05$, the DNN and XGB. Results shown are for the 36-variable networks at $\lambda = 0.05$. As in Fig. 3, the solid lines correspond to quantum (green) or simulated (blue) annealing, and dotted lines to the DNN (red) or XGB (cyan). The vertical lines denote 1σ error bars, defined by the variation over the training sets (grey) plus statistical error (green); see Supplementary Information section 6 for details of the uncertainty analysis. Whereas the DNN and XGB have an advantage for large training datasets, we find that the annealer-trained networks perform better for small training datasets. The overall performance of QAML and its features, including the advantage at small training-dataset sizes and saturation of the AUROC at approximately 0.64, are stable across a range of values of λ. An extended version of this plot, for various values of λ, is shown in Supplementary Fig. 2.

a signal efficiency ε_S we use the strong classifier sampled from the annealer with the maximum background rejection r_B. We construct such compound classifiers for simulated and quantum annealing using excited states within a fraction f of the ground-state energy E^g—that is, all $\{s_i\}$ such that $H(\{s_i\}) < (1 - f)E^g$ (note that $E^g < 0$). Simulated annealing is used as a natural comparison to quantum annealing on these fully connected problems.

In our experiments, quantum annealing struggles to find the true minimum of the objective function. This is probably a consequence of the fact that the current generation of D-Wave quantum annealers suffers from non-negligible noise on the programmed Hamiltonian. The problem of noise is compounded by the relatively sparse graph, which requires a chain of qubits to embed the fully connected logical Hamiltonian. In our case, 12 qubits are ferromagnetically coupled to act as a single logical qubit. We therefore study and interrogate current-generation quantum annealers and interpret their performance as a lower bound for the performance of future systems with lower noise and denser hardware graphs.

In Fig. 3 we plot the ROC curves illustrating the ability to discriminate between signal and background for each algorithm, with $f = 0.05$ and training datasets with 100 or 20,000 events. We observe a clear separation between the annealing-based classifiers and the binary-decision-tree-based XGB and DNN classifiers, with the advantage of the annealers appearing for small training datasets, but disappearing for the larger datasets. In Fig. 4 we plot the area under the ROC curve for each algorithm, for training datasets of various sizes and $f = 0.05$ (the largest value we used). An ideal classifier would have an area of 1. We find that quantum and simulated annealing have comparable performance, implying high robustness to approximate solutions of the training problem. This feature appears to generalize across the domain of QAML applications (Li, R. *et al*., submitted manuscript). Here the asymptotic performance of the QAML model is achieved with just 1,000 training events, and thereafter the algorithm does not benefit from additional data. This is not true for the DNN or XGB. A notable finding of our work is that QAML has an advantage over both the DNN and XGB when training datasets are small. This is shown in Fig. 5 in terms of the integral of the true negative differences over signal efficiency for various ROCs. In the same regime of small training datasets, quantum annealing develops a small advantage over simulated annealing as the fraction of excited states f used increases, saturating at $f = 0.05$. However, the uncertainties are too large to draw definitive conclusions in this regard. In the regime of large training datasets, simulated annealing has a small advantage over quantum annealing, to a significance of approximately 2σ.

In our study we have explored QAML, a simple method inspired by the prospect of using quantum annealing as an optimization technique for constructing classifiers, and applied the technique to the detection of Higgs decays. The training data are represented in a compact representation of $\mathcal{O}(N^2)$ couplers and local biases in the Hamiltonian for N weak classifiers. The resulting strong classifiers perform comparably to the state-of-the-art standard methods that are currently used in high-energy physics, and have an advantage when the training datasets are small. The role of quantum annealing is that of a subroutine for sampling the Ising problem that may in the future have advantages over classical samplers, either when used directly or as a way of seeding classical solvers with high-quality initial states.

QAML is resistant to overfitting because it involves an explicit linearization of correlations. It is also less sensitive to errors in the

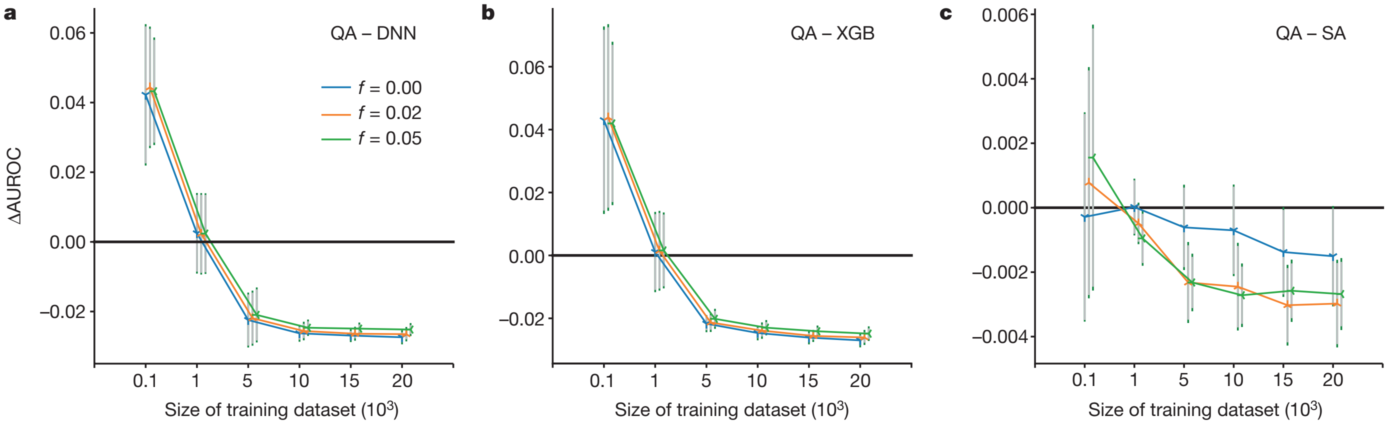

Figure 5 | Difference between the AUROCs for different algorithms. **a**, Quantum annealing versus the DNN (QA − DNN). **b**, Quantum annealing versus XGB (QA − XGB). **c**, Quantum versus simulated annealing (QA − SA). In all cases, the difference is shown as a function of training-dataset size and fraction f above the minimum energy returned (the same values of f are used for quantum and simulated annealing in **c**). Formally, we plot $\int_0^1 [r_B^{QA}(\varepsilon_S) - r_B^i(\varepsilon_S)] d\varepsilon_S$, where r_B is the maximum background rejection, $i \in \{\mathrm{DNN, XGB, SA}\}$ and ε_S is the signal efficiency. The vertical lines denote 1σ error bars. The large error bars are due to noise on the programmed Hamiltonian.

Monte Carlo correlation estimates than are DNNs or binary decision trees, owing to the truncation of the tails of the distributions (see Supplementary Information). A useful aspect of the model is that it is interpretable directly, with each weak classifier corresponding to a physically relevant variable, or product or ratio of variables, and the strong classifier being a simple linear combination thereof. This is in contrast to the creation of black-box machine learning discriminants, such as when using DNNs or XGB, developing techniques for the interpretability of which is still an active area of research[20].

Being able to use quantum annealing to optimize classifiers in a physics problem opens up further opportunities for research. There have been several recent theoretical advances in quantum machine learning[21–27]; we demonstrate that elements of this technique can already be applied to current- and next-generation quantum annealing architectures. The near future will see applications of these techniques to more complex problems in particle physics and other sciences; for example, this work has motivated similar studies in computational biology (Li, R. *et al.*, submitted manuscript). We envision that QAML could be used in the context of data certification in high-energy physics, where training with small datasets could be particularly useful. The robustness of QAML will enable both a substantial reduction in the level of human intervention and an increase in the accuracy of quickly assessing and certifying particle-collision data for analysis. Being impervious to overtraining, QAML is a good candidate for boosting[10]. We foresee studies to evaluate this application using future versions and architectures of quantum annealers or even classical optimization techniques such as simulated annealing. Multi-stage classifiers—which find the most influential variables via training and output them to form a smarter classifier—could be used to mitigate the influence of hardware noise in the quantum annealer. With more available qubits, or more efficient architectures, integer weights could be used in place of binary weights through a straightforward extension of the encoding scheme used in this work.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 April; accepted 28 July 2017.

1. Chatrchyan, S. *et al.* Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716,** 30–61 (2012).
2. Aad, G. *et al.* Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716,** 1–29 (2012).
3. Kadowaki, T. & Nishimori, H. Quantum annealing in the transverse Ising model. *Phys. Rev. E* **58,** 5355–5363 (1998).
4. Das, A. & Chakrabarti, B. K. *Colloquium:* Quantum annealing and analog quantum computation. *Rev. Mod. Phys.* **80,** 1061–1081 (2008).
5. Neven, H., Denchev, V. S., Rose, G. & Macready, W. G. Training a binary classifier with the quantum adiabatic algorithm. Preprint at http://arXiv.org/abs/0811.0416 (2008).
6. Pudenz, K. L. & Lidar, D. A. Quantum adiabatic machine learning. *Quantum Inf. Process.* **12,** 2027–2070 (2013).
7. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220,** 671–680 (1983).
8. Katzgraber, H. G., Trebst, S., Huse, D. A. & Troyer, M. Feedback-optimized parallel tempering monte carlo. *J. Stat. Mech.* **2006,** P03018 (2006).
9. Chollet, F. Keras, https://github.com/fchollet/keras (2015).
10. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. Preprint at https://arxiv.org/abs/1603.02754 (2016).
11. Khachatryan, V. *et al.* Observation of the diphoton decay of the Higgs boson and measurement of its properties. *Eur. Phys. J. C* **74,** 3076 (2014).
12. Aad, G. *et al.* Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector. *Phys. Rev. D* **90,** 112015 (2014).
13. Patrignani, C. *et al.* Review of particle physics. *Chin. Phys. C* **40,** 100001 (2016).
14. Englert, C., Plehn, T., Zerwas, D. & Zerwas, P. M. Exploring the Higgs portal. *Phys. Lett. B* **703,** 298–305 (2011).
15. Morrissey, D. E. & Ramsey-Musolf, M. J. Electroweak baryogenesis. *New J. Phys.* **14,** 125003 (2012).
16. Buttazzo, D. *et al.* Investigating the near-criticality of the Higgs boson. *J. High Energy Phys.* **12,** 89 (2013).
17. Adam-Bourdarios, C. *et al.* The Higgs machine learning challenge. *J. Phys. Conf. Ser.* **664,** 072015 (2015).
18. Albash, T., Vinci, W., Mishra, A., Warburton, P. A. & Lidar, D. A. Consistency tests of classical and quantum models for a quantum annealer. *Phys. Rev. A* **91,** 042314 (2015).
19. Al-Rfou, R. *et al.* Theano: a Python framework for fast computation of mathematical expressions. Preprint at http://arxiv.org/abs/1605.02688 (2016).
20. Chen, J. *et al.* Interpreting the prediction process of a deep network constructed from supervised topic models. In *IEEE Int. Conf. Acoustics, Speech and Signal Processing* 2429–2433 (IEEE, 2016).
21. Lloyd, S., Mohseni, M. & Rebentrost, P. Quantum principal component analysis. *Nat. Phys.* **10,** 631–633 (2014).
22. Wiebe, N., Kapoor, A. & Svore, K. Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning. *Quantum Inf. Comput.* **15,** 318–358 (2015).
23. Paparo, G. D., Dunjko, V., Makmal, A., Martin-Delgado, M. A. & Briegel, H. J. Quantum speedup for active learning agents. *Phys. Rev. X* **4,** 031002 (2014).
24. Rebentrost, P., Mohseni, M. & Lloyd, S. Quantum support vector machine for big data classification. *Phys. Rev. Lett.* **113,** 130503 (2014).
25. Schuld, M., Sinayskiy, I. & Petruccione, F. An introduction to quantum machine learning. *Contemp. Phys.* **56,** 172–185 (2015).
26. Cong, I. & Duan, L. Quantum discriminant analysis for dimensionality reduction and classification. *New J. Phys.* **18,** 073011 (2016).
27. Biamonte, J. *et al.* Quantum machine learning. *Nature* **549,** 195–202 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements This project is supported in part by the United States Department of Energy, Office of High Energy Physics Research Technology Computational HEP and Fermi Research Alliance, LLC under contract no. DE-AC02-07CH11359. The project is also supported in part under ARO grant number W911NF-12-1-0523 and NSF grant number INSPIRE-1551064. The work is supported in part by the AT&T Foundry Innovation Centers through INQNET, a programme for accelerating quantum technologies. We thank the Advanced Scientific Computing Research programme of the DOE for the opportunity to present and discuss this work at the ASCR workshop on Quantum Computing for Science (2015). We acknowledge the funding agencies and all of the scientists and staff at CERN and internationally whose hard work resulted in the momentous H(125) discovery in 2012.

Author Contributions A.M. mapped the problem on the D-Wave software architecture, analysed the data, provided the machine learning methodology and, with M.S., provided knowledge of Higgs physics. J.J. provided quantum and simulated annealing research, data analysis, machine learning work and, with D.L., quantum and simulated annealing knowledge. J.-R.V. provided quantum annealing research, data analysis, machine learning methods and error analysis, with J.J. D.L. and M.S. oversaw the work, data analysis and results. D.L. conceived the quantum machine learning methodology. M.S. conceived the application. All authors contributed to writing and reviewing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.S. (smaria@caltech.edu).

Reviewer Information *Nature* thanks S. Zohren and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Problem construction. We simulate 3×10^5 125-GeV-mass Higgs-particle decays produced by gluon fusion at $\sqrt{\hat{s}}=8$ TeV using PYTHIA 6.4[28], and 3×10^5 background events corresponding to standard-model processes using SHERPA[29]. We restrict the simulated events to those processes with realistic detector acceptance and with trigger requirements that lie directly under the Higgs peak (to ensure that the classifier cannot select on mass information); that is, those for which $|\eta|<2.5$, with one photon having $p_{\mathrm{T}}>32$ GeV and the other having $p_{\mathrm{T}}>25$ GeV, and with total diphoton invariant mass $122.5\,\mathrm{GeV}<m_{\gamma\gamma}<127.5\,\mathrm{GeV}$. The main Feynman diagrams are shown in Fig. 1. The resulting distributions for eight kinematic variables in this problem are shown in Fig. 2. The complete procedure for the construction of the weak classifier is provided in Supplementary Information.

There are 666 floating-point parameters in our Ising Hamiltonian on 36 variables. XGBoost with a maximum depth of 10 has up to 1,024 decisions (each a free variable or parameter) in each tree. Our DNN has 2,000 local biases and approximately 500,000 weights on or between the two 1,000-node hidden layers.

Data collection and analysis. For simulated annealing, all of the Ising Hamiltonians are run 10^4 times, with various numbers of linear sweeps (but all around 1,000) from an initial inverse temperature of $\beta=0.1$ to a final inverse temperature of $\beta=5$. Ground-state energies are estimated using simulated annealing with 10^4 linear sweeps from $\beta=0.1$ to $\beta=10$. For quantum annealing, we first create for each instance a heuristic embedding using the D-Wave application program interface. Quantum annealing is run with 50 gauges[18] (a randomization procedure designed to average out random errors on the local fields and couplers) at the minimum possible annealing time of 5 μs for 200 samples per gauge and a chain strength of 6.

We collect all data across programming cycles for quantum annealing and consider the ensemble of resulting solutions as a single block. For simulated and quantum annealing, we construct a histogram of the unique solutions that are returned from the algorithm, and exclude those states with low enough rates of occurrence that we cannot be certain of their inclusion in further runs. This is done by excluding any solution that occurs fewer than three times, because such solutions have a greater than 5% chance of exclusion in subsequent batches of 10^4 solutions. In this way, we determine a robust lower bound on the performance of the ensemble classifier. Details on the data and error analysis are provided in Supplementary Information.

Weak-classifier construction. We define $S(v)$ as the distribution of the signal of variable v, and similarly $B(v)$ is the distribution of the background for variable v. For a given value of v, we compute the 70th percentile of $S(v)$, v_{cut}, and then find the percentile that corresponds to $B(v_{\mathrm{cut}})$. If the percentile in $B(v_{\mathrm{cut}})$ is less than 70, then we centre v_{cut} ($v'=v-v_{\mathrm{cut}}$) and reflect across the vertical axis ($v''=-v'$) so that $S(v'')>B(v'')$ for $v''>0$, and thus the region $v''>0$ is predominantly signal and the region $v''<0$ is predominantly background. If the percentile in $B(v_{\mathrm{cut}})$ is more than 70, then we compute the 30th percentile of $S(v)$ (yielding $v_{\mathrm{cut,new}}$); if the percentile in $B(v_{\mathrm{cut,new}})$ is more than 30, then we centre at $v_{\mathrm{cut,new}}$ but do not reflect across the vertical axis, because the requirement that $S(v'')>B(v'')$ for $v''>0$ is already satisfied. If neither of these conditions is satisfied, the we reject the variable as unsuitable for the construction of a weak classifier. We then determine the 10th percentile for $S(v)$ and the 90th for $B(v)$, v_{+1} and v_{-1}, respectively. The weak classifier is

$$c(v)=\begin{cases}+1 & \text{if } v_{+1}<v''(v)\\ \dfrac{v''(v)}{v_{+1}} & \text{if } 0<v''(v)<v_{+1}\\ \dfrac{v''(v)}{|v_{-1}|} & \text{if } v_{-1}<v''(v)<0\\ -1 & \text{if } v''(v)<v_{-1}\end{cases}$$

By construction, $c(v)$ has all of the properties that we seek in a weak classifier. Because this procedure removes information about the tails of the distributions and does not take into account correlations between our kinematic variables, we introduce products and ratios of the kinematic variables to our description. If we had to flip the distribution for variable i, then we define $g_i=1/v_i$; otherwise, $g_i=v_i$. We then introduce all of the functions of the form $p(g_i,g_j)=g_ig_j$ and perform the weak-classifier construction on these combinations.

Instances and variable inclusion. We use eight kinematic variables, listed in Table 1. They involve functions of the individual and diphoton mass, as well as the angles of the photons and the diphoton system. Taking products of these kinematic variables, we obtain a total of 36 variables that pass the weak-classifier construction procedure for the vast majority of the training sets. These 36 weak classifiers (or a subset thereof) are the set from which we built our strong classifiers. For each size of training set (100, 1,000, 5,000, 10,000, 15,000 or 20,000), we generated 20 training sets and the corresponding Ising problem for $\lambda=0.05$. To compare the performance of simulated and quantum annealing, we estimate the ground-state solution of these Ising problems by running simulated annealing for a large number of sweeps (10^4) with a low final temperature (0.1 in normalized energy units).

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request. The data shown in the figures are provided as Supplementary Data.

28. Sjöstrand, T., Mrenna, S. & Skands, P. PYTHIA 6.4 physics and manual. *J. High Energy Phys.* **5,** 26 (2006).
29. Gleisberg, T. *et al.* Event generation with SHERPA 1.1. *J. High Energy Phys.* **2,** 7 (2009).