Learning interpretable representations of entanglement in quantum optics experiments using deep generative models

Daniel Flam-Shepherd, Tony C. Wu, Xuemei Gu, Alba Cervera-Lierta, Mario Krenn and Alán Aspuru-Guzik

Quantum physics experiments produce interesting phenomena such as interference or entanglement, which are the core properties of numerous future quantum technologies. The complex relationship between the setup structure of a quantum experiment and its entanglement properties is essential to fundamental research in quantum optics but is difficult to intuitively understand. We present a deep generative model of quantum optics experiments where a variational autoencoder is trained on a dataset of quantum optics experiment setups. In a series of computational experiments, we investigate the learned representation of our quantum optics variational autoencoder (QOVAE) and its internal understanding of the quantum optics world. We demonstrate that QOVAE learns an interpretable representation of quantum optics experiments and the relationship between the experiment structure and entanglement. We show QOVAE is able to generate novel experiments for highly entangled quantum states with specific distributions that match its training data. QOVAE can learn to generate specific entangled states and efficiently search the space of experiments that produce highly entangled quantum states. Importantly, we are able to interpret how QOVAE structures its latent space, finding curious patterns that we can explain in terms of quantum physics. The results demonstrate how we can use and understand the internal representations of deep generative models in a complex scientific domain. QOVAE and the insights from our investigations can be immediately applied to other physical systems.

Quantum mechanics contains a wide range of phenomena that seem counterintuitive from a classical physics perspective. Experimental quantum physics is integral to the investigation of the fundamental questions associated with these phenomena and the quantum-mechanical nature of the Universe. Quantum entanglement is one of those phenomena that is most difficult to reconcile with our picture of reality and also provides the basis for all quantum technologies and applications. Thus, in particular, quantum optics experiments are not only used to test the foundations of quantum physics but they are also at the heart of numerous quantum technologies in many areas including communication and computation. The quantum optics experiments we consider here consist of individual optical elements or devices such as lasers, beamsplitters or nonlinear crystals. Complex quantum phenomena such as multiphoton interference effects are challenging to understand intuitively. For this reason, in general, the connection between experiment structures and its entanglement properties—the so-called structure–property relation—is complicated to grasp for humans, which leads to the undiscovered potential of these technologies.

In order for the continuing advancement of fundamental research and quantum technologies, it is advantageous that researchers develop computational methods that help in the designing of new quantum hardware and providing conceptual understanding of the results. Examples include the Melvin algorithm that learns to expand its own toolbox with useful elements, or a graph-based topological optimizer that allows to extract new human-interpretable concepts. Other works show how to optimize setups with genetic algorithms, reinforcement learning or parametrized optimization. These efforts do not directly generate quantum optics experiments through the use of a learned representation trained on examples of experiments. Such an approach would provide us with the ability to generate with the prior knowledge of specific entangled experiments and allow us to directly explore the relationship between the experiment structure and entanglement in the model’s learned representation. Therefore, in this work, we focus on using deep unsupervised learning and build a generative model of quantum optics experiments.

Deep generative models have had a major impact in the past few years where they have been successfully applied to a variety of data, including images, text and audio. Generative models allow one to generate new examples similar to the training data. In particular, many advances have been made using deep generative models in the chemical sciences; for example, variational autoencoders (VAEs) have been widely used for molecular design. They enable us to generate specific distributions of molecules with certain molecular properties to efficiently search through chemical space. This offers advantages over other approaches that generate molecules without prior knowledge of the targeted distribution. In particular, VAEs allow for the efficient optimization of discrete molecular structures by learning continuous latent representations.

Learning interpretable representations of generative factors of structured data in science is an important precursor for the
development of artificial intelligence that is able to learn concepts to make scientific discoveries. For example, in a supervised setting, SciNet has been used to gain conceptual insights using time-series observations from simple physical systems; motivated by this, we investigate learning interpretable representations in an unsupervised way with structured data in a complex scientific domain.

Specifically, we study if deep generative models can learn the representations of entangled experiments in an interpretable way to efficiently explore the space of quantum optics experiments. We demonstrate that our model called quantum optics variational autoencoder (QOVAE)—a deep generative model of quantum optics experiments—can learn a representation that encodes the relationship between entanglement and experiment structure, which enables it to generate diverse and novel setups from the distributions of high-dimensionally entangled quantum optics experiments. We find that QOVAE is able to generate very specific spaces of entangled states by training on tightly constrained ranges of entanglement in experiment setups; in particular, QOVAE can generate the most challenging training experiments—with the highest entanglement and least devices—faster than it takes random sampling to produce such experiments. We further demonstrate a method using Bayesian optimization (BO) in QOVAE's latent space to search and target individual states using a novel objective that can be customized with specific constraints according to the scientist's interest. Most importantly, we show that QOVAE can learn an interpretable representation of experiments; by doing so, we discover how the model learns, opening its black box to see that QOVAE encodes experiments in its latent space according to each experiment's length and ordering of devices. Our central contribution is not solely that QOVAE can be used for experiment design, which many very powerful domain-specific methods already exist, but rather that QOVAE can autonomously learn highly complex systems in a human-interpretable way that could be understood and potentially practically exploited by scientific experts for the investigation of highly entangled quantum systems.

Quantum optics experiments
To represent each of the experiment setups, we use a discrete sequence of optical devices (Fig. 1c). Every optical device is identified by its location in the graph, specified by the photons propagating through the device and its order in the sequence. Therefore, each sequence uniquely determines the final quantum state and entanglement properties of the system. The quantum system in each experiment is a four-photon system with its initial state created by a double spontaneous parametric downconversion (SPDC) process—experiments with too few two-photon devices and specific device orderings can produce a state with a single basis ket or no entanglement at all. In total, we use the Melvin computer algorithm with a fixed device set (essentially a random search) to generate a training dataset of quantum optics experiments. This involves repeatedly random sampling experiment setups and evaluating them. To randomly sample an experiment with \( \ell \) devices, we first sample \( \ell \approx \text{uniform}[3, \ldots, T] \)—the experiment length—from a discrete uniform distribution over possible lengths from 3 to T and then sample what each device \( d \) is in the sequence \( d \approx \text{uniform}\{B_{5b}, D_{2a}, H_{2}^{0} \ldots \} \) from another discrete uniform distribution over the entire device toolbox. Next, we calculate the total entanglement \( S \) of the experiment: those with \( S>0 \) produce entangled states and ones with \( S=0 \) are unentangled. In total, we generate 200,000 entangled and unentangled setups.

Device toolbox
The experiments are generated using a set of basic elements consisting of beamsplitters (\( B_{5p} \)), mirrors (\( M \)), dove prisms (\( D \)), single-mode OAM downconverters (\( D_{5p} \)) and holograms (\( H_{2}^{0}, n \in \mathbb{Z} \)) (ref. 4). For each device, its operator is subscripted by the path(s) it acts on: either a single path \( p \) or two paths \( p \) and \( p' \). The holograms and dove prisms have discrete parameters corresponding to the OAM and phase added to the beam, respectively. We use a toolbox of six types of device operating on four possible photon paths with up to two empty paths (Fig. 1a,b). Empty paths are important for increasing the diversity of possible states. Methods provide more details on the device toolbox and how different devices change the quantum state.

Entanglement measure
The system we study is a high-dimensional four-photon quantum state. To quantify its entanglement, we derive the entanglement entropy from the discrete Schmidt rank vector (SRV)\(^{49}\). The SRV is a vector composed of the Schmidt ranks of all bipartitions which, in the case of four particles, has a size of seven. For an overall measure of entanglement, we use \( S \): the sum of all bipartition entanglement entropies, with an experiment where \( S>0 \) is entangled and \( S=0 \) is unentangled. The state and entanglement are numerically calculated using the symbolic algebra Python package SymPy (ref. 48). In general, computing the state of highly entangled experiments can be expensive; therefore, it is helpful to find more direct methods to find experiments of certain states. Towards this goal, BO in QOVAE’s latent space can be used to search for specific states.

QOVAE results
Model description. For our QOVAE model, we use a variational autoencoder to learn the distributions of quantum experiments as sequences. The QOVAE model consists of two neural networks: an encoder that maps quantum optics experiment \( x \) to continuous latent representation \( z \) and decoder that reconstructs experiments \( x \) from latent representation \( z \). Both encoder and decoder are parameterized by deep neural networks. Figure 1d displays the main model. The encoder of QOVAE learns a representation by using layers of one-dimensional convolutions that are used to generate the mean and log standard deviation of the latent space. The decoder uses the latent representation of the experiment to generate the experiment sequence using a recurrent neural network.

Setup encoding. For the training data, we sequentially represent an experiment as a series of one-hot column vectors \( x \), in a matrix where \( x = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{T \times D} \). Here \( T \) is the maximum experiment length (number of devices) and \( D \) is the number of devices in the toolbox. In any experiment, every possible device on any path or path combination is represented as a one-hot vector, namely, \( x_i \in \{0, 1\}^D \). For example, the experiment (Fig. 1e) shown as a sequence of operators, namely, \( H_{2}^{-1} \rightarrow B_{5b} \rightarrow H_{2}^{-1} \rightarrow D_{3a} \rightarrow B_{5b} \), would have five one-hot vectors for each device (and would be padded with \( T=5 \) zero vectors 0).

Training data preparation. We use the Melvin computer algorithm with a fixed device set (essentially a random search) to generate a training dataset of quantum optics experiments. This involves repeatedly random sampling experiment setups and evaluating them. To randomly sample an experiment with \( \ell \) devices, we first sample \( \ell \approx \text{uniform}[3, \ldots, T] \)—the experiment length—from a discrete uniform distribution over possible lengths from 3 to T and then sample what each device \( d \) is in the sequence \( d \approx \text{uniform}\{B_{5b}, D_{3a}, H_{2}^{-1} \ldots \} \) from another discrete uniform distribution over the entire device toolbox. Next, we calculate the total entanglement \( S \) of the experiment: those with \( S>0 \) produce entangled states and ones with \( S=0 \) are unentangled. In total, we generate 200,000 entangled and unentangled setups.

Space of entangled experiments. There is an important distinction between Haar random states and our random quantum optics experiments created by randomly assembling optical devices from a toolbox, similar to Melvin\(^{49,52,53}\). These random experiments are not guaranteed to create entangled states, based on this device toolbox—experiments with too few two-photon devices and specific device orderings can produce a state with a single basis ket or no state at all (both are \( S=0 \)). Let \( n_p \) be the number of beamsplitters or downconverters in an experiment (two-photon devices). Higher entanglement (larger \( S \)) is more likely with larger experiments and larger \( n_p \). Indeed, a
necessary but not sufficient condition for entanglement is for the experiment to satisfy $n_t > 1$. One can increase the probability of sampling two-photon devices to increase $S$, but this is challenging to balance with sampling other devices to ensure sampling diverse states. Most importantly, the experiment's device order exactly determines its entanglement and different orderings will likely produce a different entanglement $S$.

We can estimate the size of the space of entangled versus unentangled experiments through random sampling. We do this for experiments with $\ell = 6$ and $3 \leq \ell \leq 12$, sampling exactly the same as when building the training data. The results show that entangled experiments make up $33.1 \pm 4.0\%$ and $40.6 \pm 4.8\%$ of the two spaces. We report the average of five runs of sampling 10,000 setups ± standard error.

Fig. 1 | Data representation and model. 

a, Every experiment involves four photon paths beginning with an SPDC crystal and ending with a detector (represented by a grey circle with a black outline). The paths are designated by arrows and colour coded with blue, grey, red and yellow for photons a, b, c and d, respectively. In addition, there can be two empty paths e and f (green and purple arrowheads). Empty photon paths start without a crystal and are shown with a red three-pointed star and they end with no detector (the detector symbol with a slash). 

b, Visualization of the toolbox of possible devices in any experiment. 

c, Example of a quantum optics experiment, visually shown as a sequence of graphs; each device's visualization from the toolbox is shown below. 

d, Visual depiction of QOVAE. First, an experiment encoded as a sequence into a stochastic latent representation using a convolutional neural network and is reconstructed using another deep recurrent neural network. The entanglement measure is shown above as a function of the latent space. On the entanglement function, arrows are shown moving from a region of low entanglement (shown in violet) towards a region of high entanglement (shown in yellow).
Investigations. For our investigations, we train models on multiple different subsets of the compiled dataset: either QOV AE-high with a six-dimensional latent space or QOV AE-low with a two-dimensional latent space. For different investigations and datasets that we train on, we restrict the length $\ell$ of any setup and the total number of beamsplitters or downconverters, $n_p$. We conduct a number of investigations to assess QOV AE and its learned representation with different training data; in each, we generate 10,000 quantum optics experiments from trained models to investigate the entanglement properties. Further details about the model can be found in Methods. From the investigations, we extract a series of results that we discuss in the following paragraphs.

(1) QOV AE can generate novel experiments from the space of entangled or unentangled quantum optics experiments. In this investigation, we study how capable QOV AE is in generating from specific spaces of quantum optics experiment setups that produce entangled and unentangled states. We train QOV AE-high on experiments with larger lengths of $3 \leq \ell \leq 12$ and then smaller ones with $\ell = 6$ (spaces explored before).

For each length restriction, we train three different QOV AE-high models on (1) unentangled setups, (2) entangled setups and (3) a combination of the two. To ensure that QOV AE is not just counting devices to distinguish entangled and unentangled experiments, we ensure that both have a similar distribution of two-photon devices (responsible for entanglement). This is achieved by ensuring that $n_p > 1$ (that each setup has two or more two-photon devices). Supplementary Information shows the data for $\ell \leq 12$.

From these results, we find that QOV AE produces a similar percentage of experiments with $S > 0$ that exist in each dataset that it is trained on. For $\ell = 6$, when QOV AE is trained on unentangled and entangled experiments, it generates $4.4 \pm 2.3\%$ and $89.4 \pm 3.3\%$ $S > 0$, respectively, as well as $50.7 \pm 1.2\%$ when trained on a $48.1\%$ combination of $S > 0$ and $S = 0$. For $\ell \leq 12$, when QOV AE is trained on unentangled and entangled experiments, it generates $7.3 \pm 4.3\%$ and $91.5 \pm 2.6\%$ $S > 0$, respectively, as well as $52.3 \pm 5.5\%$ when trained on a $50.9 \pm 0.0\%$ mixture of $S > 0$ and $S = 0$. We report the average percentage of five sets of generated experiments $\pm$ standard error. For the entangled data, the percentage of $S > 0$ achieved is comparable to the validity scores of generative models of molecules.52

For these models (and subsequent investigations), we also find that QOV AE generates 99\% unique experiments, essentially producing no duplicate experiments. Furthermore, it generates all the novel experiments that do not appear in the training data.

Based on these results, we observe that QOV AE can learn to generate from the space of quantum optics experiments with different entanglement properties.

(2) QOV AE can learn distributions of entangled states that it is trained on. Figure 2a,b displays 15 random samples of experiments from QOV AE-high and its training setups with lengths $\ell \leq 12$ and $0 < S \leq 7.5$. We can see that the model has learned to generate experiments that have a similar structure to the training experiments as both sets of samples have similar numbers of one- and two-photon devices, as well as empty path devices.

We test if the distribution of entanglement of every bipartition is similar between experiments from the training data and the model. We generate experiments from the model and take training data and calculate all their entanglement entropies and Schmidt ranks in all the seven bipartitions. We visually compare the distributions with Fig. 2 | Comparison with training data. a. Distribution plots for entanglement entropy and the Schmidt rank of all the seven system bipartitions calculated using experiments generated by QOVAE-high and training experiments. b, c. Training experiments (b) and experiments generated (c) by QOVAE-high.
Articles

We encode two experiments with $S$ = 2.77 and another with $S$ = 5.55. Each level of entanglement basically defines a group of states in these groups.

We report that QOV AE can learn precise distributions of different levels of entanglement, defined by small intervals of entanglement with unity length. We train QOV AE on four datasets with entanglement levels of (1) $2.0 < S < 3.0$, (2) $3.0 < S < 4.0$, (3) $4.0 < S < 5.0$ and (4) $5.0 < S < 6.0$ (using 14,000, 11,000, 18,000 and 10,000 setups, respectively). For each level of $S$, the results in Table 1 show that QOV AE perfectly matches the training mode of $S$ (the $S$ value most likely to be sampled) and learns the training mean $S$ within one standard error but overestimates the standard deviation of $S$, confirming that QOV AE seems to learn heavier right tails for $S$ (Fig. 2a). The model can perfectly capture the mode with no error. This is expected for well-trained QOV AE, whereas when $S$ is continuous, all the training setups do not have different values; in fact, one prominent value of $S$ dominates. For example, in the $4.0 < S < 5.0$ level, there are 677 different $S$ values in the 18,000 training setups but 40% of the 18,000 training setups have $S$ = 4.39. Each level of entanglement basically defines a group of similar states; therefore, QOV AE can learn to target and generate states in these groups.

(4) **QOV AE can be used to efficiently search for new highly entangled experiments and states.** We find that QOV AE can efficiently search for new experiments with higher entanglement than random search and that QOV AE is able to target and generate from specific levels of entanglement as well as it can generate the most entangled training experiments with fewest devices, far faster than the random methods that produced them. Additionally, BO can be used in QOV AE’s latent space to exactly target specific states.

We show three latent space interpolations (Fig. 3) from different entanglement measures $S$. The first interpolation (Fig. 3) interpolates between experiments that do not produce entangled states; we encode two experiments with $S$ = 0 and generate experiments along the latent path between them. In this interpolation, the experiments decoded along the interpolation path remain in the unentangled ($S$ = 0) space like the initial and final experiments.

The second is between experiments that both have $S$ = 5.55 and the last between one experiment with $S$ = 2.77 and another with $S$ = 5.55. In the second interpolation, the decoded experiment entanglement is within 0.50 of the initial and final experiments ($S$ = 5.50). In the last interpolation, the entanglement $S$ of the setups decoded along the path increases linearly from $S$ = 2.77 towards the final experiment’s entanglement of $S$ = 5.55.

We also evaluate if nearby experiments in the latent space have similar entanglement properties by comparing the Euclidean distance between the latent points and their absolute entanglement difference. We find that as the latent distance increases, the absolute entanglement difference increases and when the latent distance goes to 0, the difference in entanglement does too. A plot of this relationship obtained by random sampling is shown in the Supplementary Information. Thus, nearby experiments in the latent space are more likely to have similar entanglement properties than experiments further away.

Hence, QOV AE learns a representation that encodes a measure of similarity between the experiment and entanglement $S$.

(3) **QOV AE learns a quasi-continuous embedding in terms of entanglement.** We demonstrate the smoothness of the latent space in terms of entanglement $S$ by testing if experiments that are close in the latent space have similar entanglement. First, we perform spherical interpolations\(^4\) from one latent representation $z_i$ of an experiment to another $z_j$, decoding four experiments, at equally spaced steps on the interpolated path from $z_i$ to $z_j$.

We show three latent space interpolations (Fig. 3) from different entanglement measures $S$. The first interpolation (Fig. 3) interpolates between experiments that do not produce entangled states; we encode two experiments with $S$ = 0 and generate experiments along the latent path between them. In this interpolation, the experiments decoded along the interpolation path remain in the unentangled ($S$ = 0) space like the initial and final experiments.

The second is between experiments that both have $S$ = 5.55 and the last between one experiment with $S$ = 2.77 and another with $S$ = 5.55. In the second interpolation, the decoded experiment entanglement is within 0.50 of the initial and final experiments ($S$ = 5.50). In the last interpolation, the entanglement $S$ of the setups decoded along the path increases linearly from $S$ = 2.77 towards the final experiment’s entanglement of $S$ = 5.55.

We also evaluate if nearby experiments in the latent space have similar entanglement properties by comparing the Euclidean distance between the latent points and their absolute entanglement difference. We find that as the latent distance increases, the absolute entanglement difference increases and when the latent distance goes to 0, the difference in entanglement does too. A plot of this relationship obtained by random sampling is shown in the Supplementary Information. Thus, nearby experiments in the latent space are more likely to have similar entanglement properties than experiments further away.

Hence, QOV AE learns a representation that encodes a measure of similarity between the experiment and entanglement $S$.

(4) **QOV AE can be used to efficiently search for new highly entangled experiments and states.** We find that QOV AE can efficiently search for new experiments with higher entanglement than random search and that QOV AE is able to target and generate from specific levels of entanglement as well as it can generate the most entangled training experiments with fewest devices, far faster than the random methods that produced them. Additionally, BO can be used in QOV AE’s latent space to exactly target specific states.

We report that QOV AE can learn precise distributions of different levels of entanglement, defined by small intervals of entanglement with unity length. We train QOV AE on four datasets with entanglement levels of (1) $2.0 < S < 3.0$, (2) $3.0 < S < 4.0$, (3) $4.0 < S < 5.0$ and (4) $5.0 < S < 6.0$ (using 14,000, 11,000, 18,000 and 10,000 setups, respectively). For each level of $S$, the results in Table 1 show that QOV AE perfectly matches the training mode of $S$ (the $S$ value most likely to be sampled) and learns the training mean $S$ within one standard error but overestimates the standard deviation of $S$, confirming that QOV AE seems to learn heavier right tails for $S$ (Fig. 2a). The model can perfectly capture the mode with no error. This is expected for well-trained QOV AE, whereas when $S$ is continuous, all the training setups do not have different values; in fact, one prominent value of $S$ dominates. For example, in the $4.0 < S < 5.0$ level, there are 677 different $S$ values in the 18,000 training setups but 40% of the 18,000 training setups have $S$ = 4.39. Each level of entanglement basically defines a group of similar states; therefore, QOV AE can learn to target and generate states in these groups.

**Fig. 3 | Interpolations.** Three examples of latent space interpolations from QOVAE. Each interpolation is between experiments with different entanglement measure $S$. The chosen initial and final experiment along the interpolation path are enclosed in a rectangle and four experiments along the path in between the two are displayed. Below each experiment, the entanglement measure $S$ is shown.
Next, we train QOVAE-high on the most difficult training experiments to find the following: almost 5,000 from a total of 400,000 that have the highest entanglement ($S > 8.0$) and simplest structure (ten or fewer devices). The results show that QOVAE can learn to directly generate from this space as almost all the 85.4 ± 3.64% of the generated experiments have $S > 8.0$ (and ten or fewer devices). It also learns to produce the same mode ($S = 8.79$) as the training data and within one standard error of the mean entanglement (8.55 ± 0.35% for $S = 8.84$). Most importantly, using QOVAE, we can generate almost twice the number of these experiments with the highest training entanglement and least number of devices, at a small fraction of the time (hours instead of days) it took to produce these experiments when building the training dataset. This demonstrates an important advantage of QOVAE’s prior knowledge and is especially useful given how expensive entanglement is to calculate.

We can also target specific states by performing BO in the latent space; taking inspiration from other studies, we use the target objective defined as $y(s) = |\langle \psi_s | \psi_s \rangle|^2 - \text{length}(s)/4d$. This is the fidelity between the target state $|\psi_s\rangle$ and state $|\psi_s\rangle$ of experiment setup $s$ penalized by experiment length ($d$ is the maximum experiment length) and $\lambda$ is a parameter that can be tuned to strengthen or weaken the device penalty. To perform BO, we first train QOVAE so that each training experiment has a latent vector defined by the mean of the encoder. Afterwards, we train a sparse Gaussian process to predict $y(s)$ given its latent representation. Then, we perform a set number of iterations of batched BO using the expected improvement heuristic, compiling the top scoring states.

As an example, we target the two-dimensional four-photon Greenberger–Horne–Zeilinger state as

$$|\psi_s\rangle = \frac{1}{\sqrt{6}} (|0000\rangle + |1111\rangle),$$

which is the state where each bipartition has a Schmidt rank of 2 and entropy of 0.693 for a total entanglement measure of $S = 4.852$. We set the device penalty to be very small using $\lambda = 0.1$ and run BO for five iterations. We find a single quantum optics experiment that has the exact targeted state $((|0000\rangle |1111\rangle)^2 = 1$. This experiment has a total of 12 devices: three reflection devices, five beamsplitters, five holograms and one dove prism that are in the following sequence:

$$R_a \rightarrow H_d^{-1} \rightarrow BS_{bc} \rightarrow D_d \rightarrow R_c \rightarrow H_b^{-1} \rightarrow R_d \rightarrow BS_{sh} \rightarrow BS_{cd} \rightarrow BS_{ac} \rightarrow BS_{ac}. $$

This demonstrates an important advantage of QOVAE’s prior knowledge and is especially useful given how expensive entanglement is to calculate.

We can also target specific states by performing BO in the latent space; taking inspiration from other studies, we use the target objective defined as $y(s) = |\langle \psi_s | \psi_s \rangle|^2 - \text{length}(s)/4d$. This is the fidelity between the target state $|\psi_s\rangle$ and state $|\psi_s\rangle$ of experiment setup $s$ penalized by experiment length ($d$ is the maximum experiment length) and $\lambda$ is a parameter that can be tuned to strengthen or weaken the device penalty. To perform BO, we first train QOVAE so that each training experiment has a latent vector defined by the mean of the encoder. Afterwards, we train a sparse Gaussian process to predict $y(s)$ given its latent representation. Then, we perform a set number of iterations of batched BO using the expected improvement heuristic, compiling the top scoring states.

As an example, we target the two-dimensional four-photon Greenberger–Horne–Zeilinger state as

$$|\psi_s\rangle = \frac{1}{\sqrt{6}} (|0000\rangle + |1111\rangle),$$

which is the state where each bipartition has a Schmidt rank of 2 and entropy of 0.693 for a total entanglement measure of $S = 4.852$. We set the device penalty to be very small using $\lambda = 0.1$ and run BO for five iterations. We find a single quantum optics experiment that has the exact targeted state $((|0000\rangle |1111\rangle)^2 = 1$. This experiment has a total of 12 devices: three reflection devices, five beamsplitters, five holograms and one dove prism that are in the following sequence:

$$R_a \rightarrow H_d^{-1} \rightarrow BS_{bc} \rightarrow D_d \rightarrow R_c \rightarrow H_b^{-1} \rightarrow R_d \rightarrow BS_{sh} \rightarrow BS_{cd} \rightarrow BS_{ac} \rightarrow BS_{ac}. $$

This also explains why regions B and C have different entanglement $S$: in region C, the last element is $DC_{ac}$ that is needed to create entanglement, whereas in region B, the last device contributes nothing to entanglement, so these experiments are less likely to be entangled.

We can further cluster all the possible last elements into eight functional groups, for instance, all the holograms have the same effect on the entanglement property, so we can combine them into a single functional group. We plot the colour-coded latent space according to functional groups (Fig. 4d) and see additional grouping.

We conclude that QOVAE learns to structure its latent space according to the experiment length and final device.

It seems that there was pronounced correlation between the length and entanglement in the last training subset of experiments. We train another QOVAE-low on entirely entangled states and show that this is not the defining structure that QOVAE uses to learn. This time, we use 18,000 experiments evenly split between low entanglement ($3.0 < S < 4.0$) and high entanglement ($8.0 < S < 9.5$). Similarly, in Fig. 5, it is apparent that QOVAE structured its latent space by length (Fig. 5b) and last device type (Fig. 5b) as before; however, there are clearer patterns and groups when plotting the functionality of the last device type (Fig. 5d).
However, here we find that when we plot the entanglement in latent space (Fig. 5a), this does not correlate with the length of the experiments (Fig. 5b) or the last devices (Fig. 5c). Interestingly, there are visible regions of low entanglement and high entanglement in the latent space that is thus unexplained by structure. Probably, QOV AE is learning some additional information about the sequence structure or order that is responsible for this.

We further investigate how QOV AE structures its latent space by removing length from the equation such that the model cannot use it so that we can have additional insights into how QOV AE structures its latent space and encodes the experiment sequence. We do this by training on 16,000 experiments that have exactly ten devices.

We discover that QOV AE does not structure its latent space using only the last device but will also use other information and devices in the experiment sequence; one clear example of this phenomenon is the latent space shown in Fig. 5e,f. Here QOV AE-low uses not just the last device but also the second-last device to structure its latent space and group them in rows that are perpendicular to each other. It is clear that QOV AE is using the dimensions of its latent space to store information about specific devices in the sequence ordering. QOV AE-low uses its two dimensions to encode the last and second-last experiment devices in groups in the latent space. This means that QOV AE's with higher-dimensional latent spaces will be able to store more information about the devices at each location in the sequence ordering using additional dimensions to encode and group devices.

This description provides us with a complete interpretation of our model’s latent representation and how it learns. With this new understanding, we can also explain results that we have observed in previous investigations. For example, we can explain exactly how QOV AE can learn to generate from distributions of highly entangled states: it simply learns the distribution of device orderings that define its training experiments.

This distinguishes QOV AE from other generative models applied to scientific domains; by directly training on the device sequence, we enable QOV AE to learn a human-interpretable representation that could be practically exploited to design experiments. In doing so, we provide one of the very few examples of opening up and successfully demystifying black-box deep generative models applied to complex scientific systems.

Conclusions

We presented QOV AE—a deep generative model—for the design of quantum optics hardware. Deep generative models are widely used, but there has never been any investigation or understanding developed of their internal representation in a complex scientific domain. In a series of complex computational experiments, we investigated QOV AE’s internal picture of the quantum world. QOV AE was able

---

**Fig. 4 | QOV AE latent spaces.** a–d, Two-dimensional latent space of QOV AE-low trained on equal numbers of $S=0$ and $0<S<7$ with $\ell<15$. Latent points coloured by the entanglement measure $S$ (a), length (b), last devices (c) and functionality of the last device (d).
to generate novel entangled experiments, learn distributions of entanglement and was shown to smoothly interpolate in its latent space, which can also be used to efficiently search for new highly entangled experiments and target specific distributions of entangled states. When plotting QOVÆ’s latent space, we find complex internal structures where QOVÆ learns a latent representation

Fig. 5 | More QOVÆ latent spaces. a–d, Latent spaces from QOVÆ trained on 18,000 experiments with $3.0 < S < 4.0$ and $8.0 < S < 9.5$. e,f, Latent spaces from QOVÆ trained on 16,000 experiments with ten devices. Latent points coloured by the entanglement measure $S$ (a), length (b), last devices (c), functionality of last device (d), type of second-last device (e) and type of last device (f).
based on the experiment length and device ordering. Our results go beyond designing new quantum optics; they tackle the question of interpretability and explainability of black-box models in a scientific domain. This is particularly promising since understanding what these models learn could lead to new computer-inspired scientific insights and discoveries.

**Optical devices.** To ensure that the device toolbox is suitable, all the elements are used in standard quantum optics laboratories, including the technique of entanglement by path identity. In addition, we use a device toolbox that is similar to others that have been used in related work on machine-learning-assisted design of quantum optics experiments.

In our current approach, we use linear optics devices; however, it would be straightforward to include any other nonlinear optics devices like four-wave mixers or optical parametric oscillators. In addition, it would be possible to train QOV AE on a universal set of optical devices with continuous parameters by discretizing the space of those parameters. However, the performance of QOV AE needs to be investigated when including such devices. Furthermore, QOV AE could still produce experiments that are challenging to implement in the lab.

**Future work.** In this paper, we provide an initial demonstration of QOV AE and its potential; however, much further work needs to be done to truly assess and explore the use of its interpretable representation to investigate quantum optics experiments. Furthermore, although these investigations are exploratory, this work paves the way for many promising extensions that can be considered: conditional VAEs, where the encoder and decoder would be trained on both experiment and entanglement or other information about the quantum state to conditionally generate experiments. Another interesting extension is the hierarchical VAE, which has layers of latent variables; these models have the ability to learn hierarchical features of the training data and have the potential to encode more information about the experiment structure into their learned representation. Using BO in QOV AE’s latent space, it is possible to target many specific states, but this is unlikely to be as efficient as domain-specific methods. More interesting is how QOV AE learns to construct its latent representation using its latent dimensions to independently encode the ordering of the experiment device sequence. Going forward, the most important future work is to figure out how this phenomenon can be exploited to develop new insights about quantum optics experiments and their design. Additionally, in future work, searching QOV AE’s latent space, different metrics may be useful for different states, for example, like the Greenberger–Horne–Zeilinger state or W states or any other specific experiment setups for high-dimensionally entangled states. In place of fidelity $d_{FO} = \left(\sum \sqrt{p_i q_i}\right)^2$, the mean squared error between the states can be used: $d_{SE} = \sqrt{\sum (p_i - q_i)^2}$, where $p_i$ and $q_i$ are the basis ket probabilities of the target and evaluated states, respectively. Another potential metric is the Kullback–Leibler (KL) divergence between states, namely, $d_{KL} = \sum p_i \log \left(\frac{p_i}{q_i}\right)$. It is also possible to target absolutely maximally entangled states using the entanglement measure or the sum of the SRV components in place of fidelity.

**Other applications.** Indeed, the insights from this work are applicable to generative models of molecules. Molecules and quantum optics experiments are similar discrete, structured objects, representable as graphs or sequences. If we treat quantum optics experiments as undirected, we can even plot them as molecules by mapping devices to atoms (Supplementary Information provides a few examples). Based on how QOV AE learns its representation, we can infer that ChemVAE could learn a representation that is more human interpretable by using its dimensions to store information about the ordering of the SMILES string. QOV AE could, in principle, be directly applied to other physical science domains, such as the design of new quantum circuits for quantum computing. Currently, noisy intermediate-scale quantum computing algorithms are promising candidates that can surpass the classical computational capabilities for numerous applications. Most of these approaches require good priors to efficiently explore and represent the space of parameters and solutions. The exponentially large Hilbert space formed by all the possible quantum circuits makes this task computationally intractable when the structure–property relation of these circuits is still not fully understood. Most attempts to search this space have been made using genetic algorithms, which lack a sufficient prior. QOV AE’s ability to learn meaningful representations as understood by domain experts could provide insights into how the Hilbert space is organized within these parameterized quantum circuits. QOV AE learns an interpretable representation of entanglement in quantum optics experiments. Our work with QOV AE is an example in the physical sciences of opening the black box of deep generative models to develop promising scientific insights.

**Methods.**

**Encoder.** Our encoder is a diagonal Gaussian whose parameters are a mapping from the data manifold to the latent space. Our data from the quantum optics experiments $x \in \mathbb{R}^{n \times T}$ are represented as a sequence with $T$ elements, each from a toolbox of $d$ possible devices:

$$q_d(x) = N(x|\mu_d(x), \sigma_d^2(x)),$$

where $\mu, \log \sigma = q_d(x)$. First, $g$ consists of a convolutional neural network with three layers:

$$h = \text{Conv1d}_3(\text{Conv1d}_2(\text{Conv1d}_1(x))).$$

A single layer takes the form

$$x' = \text{Conv1d}(x) = \text{ReLU}(w \odot x + b),$$

where $w \in \mathbb{R}^{t \times T \times d}$ is the convolution filter tensor consisting of $n$ filters each with length $t$ and $d$ features. Also, the layer output is $x' \in \mathbb{R}^{n \times T - t + 1}$. For the input layer, this is just $d = D$, the number of devices in the toolbox used to create any experiment. ReLU() is the element-wise rectified linear unit function and $\otimes$ is the convolution operator that outputs a tensor with the following $f$ elements, each from a convolution filter $w_f$:

$$x = \sum_{f=1}^{F} w_f \odot x_{s+1}.$$

The second component of the encoder ($g$) is a multilayer perceptron (MLP) with three layers that maps the flattened output $h$ from the convolutional neural net to the parameters of latent distribution:

$$\mu, \log \sigma = \text{MLP}_\theta(\text{Flatten}(h)).$$

**Decoder.** For the observation model, every data point is a sequence of devices from the $d$ element toolbox of possible device elements; thus, we can model the data as independent categoricals whose mean vector is mapped from latent samples using a neural network:

$$p_d(x|z) = \prod_{l=1}^{T} \text{Categorical}(x_l),$$

where $p_{z_i} \in [0, 1]^d$ is the probability vector of each device in the toolbox. Our encoder outputs these probabilities as

$$p_{z_1}, \ldots, p_{z_t}, \ldots, p_{z_T} = f(z).$$

First, $f$ consists of a three-layer recurrent neural network each with gated recurrent units (GRUs):

$$h = \text{GRU}_1(\text{GRU}_2(\text{GRU}_3(\text{MLP}(z)))),$$

where $\text{MLP}_\theta$ is a multilayer perceptron (MLP) with three layers that maps the flattened output $h$ from the convolutional neural net to the parameters of latent distribution:

$$\mu, \log \sigma = \text{MLP}_\theta(\text{Flatten}(h)).$$

$$h = \text{GRU}_1(\text{GRU}_2(\text{GRU}_3(\text{MLP}(z)))),$$

where $\text{MLP}_\theta$ is a multilayer perceptron (MLP) with three layers that maps the flattened output $h$ from the convolutional neural net to the parameters of latent distribution:

$$\mu, \log \sigma = \text{MLP}_\theta(\text{Flatten}(h)).$$
where MLP is a single-layer MLP with ReLU(·) activation. A single-layer GRU takes the following form:

\[ z_t' = \sigma (W_s x_t + U_h h_{t-1} + b), \]
\[ r_t = \sigma (W_r x_t + U_h h_{t-1} + b), \]
\[ h_t' = \tanh (W_x x_t + U_h (r_t \circ h_{t-1}) + b), \]
\[ h_t = (1 - z_t') \circ h_{t-1} + z_t' \circ h_t', \]

where W, U, and b are parameters of the GRU layer. Furthermore, z_t' and r_t are update and reset gate vectors, respectively, and \( \sigma \) is the sigmoid function. The input to the first GRU layer is the t-length sequence of vectors (MLP(x), ...MLP(z)). The output of the recurrent neural network is mapped to the device probabilities using the following softmax layer:

\[ p_i = \text{Softmax}(W_h + b) = 1, \ldots, T. \]

Training and data details. We manually performed an initial hyperparameter search for QOVAE. We found an initial set of training hyperparameters using grid search. We found training on around 1,600 epochs to produce better validation accuracy and evidence of the value lower bound. We trained QOVAE using stochastic gradient descent and Adam optimizer with a low learning rate of \( \sim 10^{-4} \). Training was done using the KERAS machine-learning package from TensorFlow. All the models were trained on a V100 graphics processing unit node on the Béluga supercomputer. The minibatch size grid was \{32, 64, 128, 256\} with an optimal batch size of 64. The model architecture is the same for all the experiments with both models. We use grid search to search over the encoder and decoder architecture. For the convolutional layers, we search over 6, 12, 18 and decoder architecture. For the convolutional layers, we search over 6, 12, 18, 24, and 36 filters with lengths of 3, 4 and 5. For the MLPs in both encoder and decoder, we consider 32, 64, 128 and 256 hidden units in each layer. For the GRU layer, we consider hidden states of size 128 and a 64-unit layer that maps to the GRU layers. For new investigations, researchers can use the dataset provided to train QOVAE, which has both \( S > 0 \) and \( S = 0 \) experiments. However, they can also produce new training datasets using the Melvin computer algorithm with a different device toolbox of interest. For this, SymPy symbolic algebra is required to specify how each device changes the state of the system. After training QOVAE, they can investigate the learned representation by plotting the latent vectors for each training experiment in two or three-dimensional space. If QOVAE has a higher-dimensional latent space, they can plot specific dimensions or project the space into two or three dimensions. Based on this, the researcher can investigate how the latent space represents the training entanglement or use it to search for specific states as they see fit, either with random search or another search algorithm like the BO approach used here.

Data availability
The training data are available via GitHub at https://github.com/danielflamshep/qovae/blob/main/setup.smi (ref. 31). The source code for the Melvin algorithm is available via GitHub at https://github.com/XuemeiGu/MelvinPython.

Received: 4 October 2021; Accepted: 1 May 2022; Published online: 16 June 2022

References
1. Schrödinger, E. Discussion of probability relations between separated systems. Math. Proc. Camb. Philos. Soc. 31, 555–563 (1935).
2. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? Phys. Rev. 47, 777 (1935).
3. Bell, J. S. On the Einstein Podolsky Rosen paradox. Phys. Rev. Lett. 115, 250402 (2015).
4. Bong, K.-W. et al. A strong no-go theorem on the Wigner’s friend paradox. Nat. Phys. 16, 1199–1205 (2020).
5. Yin, J. et al. Satellite-to-ground entanglement-based quantum key distribution. Phys. Rev. Lett. 119, 208501 (2017).
6. Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. Nat. Commun. 5, 4213 (2014).
7. Paesani, S. et al. Generation and sampling of quantum states of light in a silicon chip. Nat. Phys. 15, 925–929 (2019).
8. Schrödinger, E. Discussion of probability relations between separated systems. Math. Proc. Camb. Philos. Soc. 31, 555–563 (1935).
9. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? Phys. Rev. 47, 777 (1935).
10. Bell, J. S. On the Einstein Podolsky Rosen paradox. Phys. Rev. Lett. 115, 250402 (2015).
11. Bong, K.-W. et al. A strong no-go theorem on the Wigner’s friend paradox. Nat. Phys. 16, 1199–1205 (2020).
12. Yin, J. et al. Satellite-to-ground entanglement-based quantum key distribution. Phys. Rev. Lett. 119, 208501 (2017).
13. Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. Nat. Commun. 5, 4213 (2014).
14. Paesani, S. et al. Generation and sampling of quantum states of light in a silicon chip. Nat. Phys. 15, 925–929 (2019).
15. Schrödinger, E. Discussion of probability relations between separated systems. Math. Proc. Camb. Philos. Soc. 31, 555–563 (1935).
16. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? Phys. Rev. 47, 777 (1935).
17. Bell, J. S. On the Einstein Podolsky Rosen paradox. Phys. Rev. Lett. 115, 250402 (2015).
18. Bong, K.-W. et al. A strong no-go theorem on the Wigner’s friend paradox. Nat. Phys. 16, 1199–1205 (2020).
19. Yin, J. et al. Satellite-to-ground entanglement-based quantum key distribution. Phys. Rev. Lett. 119, 208501 (2017).
20. Peruzzo, A. et al. A variational eigenvalue solver on a photonic quantum processor. Nat. Commun. 5, 4213 (2014).
21. Paesani, S. et al. Generation and sampling of quantum states of light in a silicon chip. Nat. Phys. 15, 925–929 (2019).
42. Allen, L., Beijersbergen, M. W., Spreeuw, R. J. C. & Woerdman, J. P. Orbital angular momentum of light and the transformation of Laguerre–Gaussian laser modes. Phys. Rev. A 45, 8185–8189 (1992).

43. Romero, J., Giovannini, D., Franke-Arnold, S., Barnett, S. M. & Padgett, M. J. Increasing the dimension in high-dimensional two-photon orbital angular momentum entanglement. Phys. Rev. A 86, 012334 (2012).

44. Krenn, M. et al. Generation and confirmation of a (100 x 100)-dimensional entangled quantum system. Proc. Natl Acad. Sci. USA 111, 6243–6247 (2014).

45. Erhard, M., Malik, M., Krenn, M. & Zeilinger, A. Experimental Greenberger–Horne–Zeilinger entanglement beyond qubits. Nat. Photon. 12, 759–764 (2018).

46. Y.-H., L. et al. Quantum teleportation in high dimensions. Phys. Rev. Lett. 123, 070505 (2019).

47. Leach, J., Padgett, M. J., Barnett, S. M., Franke-Arnold, S. & Courtial, J. Measuring the orbital angular momentum of a single photon. Phys. Rev. Lett. 88, 257901 (2002).

48. Huber, M. & De Vicente, J. J. Structure of multidimensional entanglement in multipartite systems. Phys. Rev. Lett. 110, 030501 (2013).

49. Meurer, A. et al. SymPy: symbolic computing in Python. Peerj. Comput. Sci. 3, e103 (2017).

50. Hamma, A., Santra, S. & Zanardi, P. Quantum entanglement in random physical states. Phys. Rev. Lett. 109, 040502 (2012).

51. Adler, T. et al. Quantum optical experiments modeled by long short-term memory. Photonics 8, 535 (2021).

52. Li, Y., Vinyls, O., Dyer, C., Pascual, R. & Battaglia, P. Learning deep generative models of graphs. In Proc. 35th International Conference on Machine Learning 1–22 (PMLR, 2018).

53. Scott, D. W. Multivariate Density Estimation: Theory, Practice, and Visualization (John Wiley & Sons, 2015).

54. Shoemaker, K. Animating rotation with quaternion curves. In Proc. 12th Annual Conference on Computer Graphics and Interactive Techniques 245–254 (ACM, 1985).

55. Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. Grammar variational autoencoder. In Proc. 34th International Conference on Machine Learning 1945–1954 (PMLR, 2017).

56. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215 (2019).

57. Gilpin, L. H. et al. Explaining explanations: an overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) 80–89 (IEEE, 2018).

58. Erhard, M., Krenn, M. & Zeilinger, A. Advances in high-dimensional quantum entanglement. Nat. Rev. Phys. 2, 365–381 (2020).

59. Pan, J.-W. et al. Multiphoton entanglement and interferometry. Rev. Mod. Phys. 84, 777 (2012).

60. Krenn, M., Hochrainer, A., Lahiri, M. & Zeilinger, A. Entanglement by path identity. Phys. Rev. Lett. 118, 080401 (2017).

61. Malik, M. et al. Multi-photon entanglement in high dimensions. Nat. Photon. 10, 248–252 (2016).

62. Wang, L., Hong, C. & Friberg, S. Generation of correlated photons via four-wave mixing in optical fibres. J. Opt. B 3, 346 (2001).

63. Giordmaine, J. A. & Miller, R. C. Tunable coherent parametric oscillation in LiNbO3, at optical frequencies. Phys. Rev. Lett. 14, 973 (1965).

64. Kingma, D. P., Mohamed, S., Rezende, D. J. & Welling, M. Semi-supervised learning with deep generative models. In Proc. 30th International Conference on Machine Learning 270–277 (JMLR, 2013).

65. Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K. & Winther, O. Ladder variational autoencoders. In Advances in Neural Information Processing Systems 1–9 (NIPS, 2016).

66. Zhao, S., Song, J. & Ermon, S. Learning hierarchical features from generative models. In Proc. 34th International Conference on Machine Learning 4091–4099 (PMLR, 2017).

67. Dür, W., Vidal, G. & Cirac, J. I. Three qubits can be entangled in two inequivalent ways. Phys. Rev. A 62, 062314 (2000).

68. Cervera-Lierta, A., Latorre, J. I. & Goyeneche, D. Quantum circuits for maximally entangled states. Phys. Rev. A 100, 022342 (2019).

69. Helwig, W. and Cui, W. Absolutley maximally entangled states: existence and applications. Preprint at https://arxiv.org/abs/1306.2536 (2013).

70. Weinger.D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36 (1988).

71. Preskill, J. Quantum computing in the NISQ era and beyond. Quantum 2, 79 (2018).

72. Cerezo, M. et al. Variational quantum algorithms. Nat. Rev. Phys. 3, 625–644 (2021).

73. Bharti, K. et al. Noisy intermediate-scale quantum (NISQ) algorithms. Rev. Mod. Phys. 94, 15004–15073 (2022).

74. Giraldi, G. A. Portugal, R. & Thess, R. N. Genetic algorithms and quantum computation. Preprint at https://arxiv.org/abs/cs/0403003 (2004).

75. Yabuki, T. & Iba, H. Genetic algorithms for quantum circuit design—evolving a simpler teleportation circuit. In Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference 421–425 (Citeseer, 2000).

76. Anand, A., Degrooto, M. & Aspuru-Guzik, A. Natural evolutionary strategies for variational quantum computation. Mach. Learn. Sci. Technol. 2, 045012 (2021).

77. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. J. Mach. Learn. Res. 14, 1303–1347 (2013).

78. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In 12th Symposium on Operating Systems Design and Implementation (OSDI 16) 265–283 (USENIX, 2016).

79. Flam-Shepherd, D. Python code and data for training and sampling from models. Zenodo https://doi.org/10.5281/zenodo.6499886 (2021).

80. Nogueira, F. Bayesian optimization: open source constrained global optimization tool for Python. GitHub https://github.com/fmfn/BayesianOptimization (2014).

Acknowledgements
A.A.-G. acknowledges support from the Canada 150 Research Chairs Program, the Canada Industrial Research Chair Program and from Google in the form of a Google Focused Award. M.K. acknowledges support from the FWF (Austrian Science Fund) via Erwin Schrödinger Fellowship no. J4309.

Author contributions
D.F.-S. conceived the overall project, developed the approach and wrote the paper. D.F.-S. and T.W. designed and performed the investigations. A.C.-L. provided the technical advice. X.G. provided the technical advice and wrote the entanglement calculation code. M.K. built the dataset, provided technical advice and helped design the interpretability investigation and analysed the experiments. A.A.-G. led the project and provided the overall directions. All the authors participated in preparing the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00493-5.

Correspondence and requests for materials should be addressed to Daniel Flam-Shepherd, Mario Krenn or Alan Aspuru-Guzik.

Peer review information Nature Machine Intelligence thanks Michael Hartmann, Sohaib Alam and Patrick Huembeli for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022.