Cambridge Quantum

Variational Inference with a Quantum Computer

by Marcello Benedetti, Brian Coyle, Mattia Fiorentini, Michael Lubasch, Matthias Rosenkranz

Cambridge Quantum
MARCELLO BENEDETTI
marcello.benedetti@cambridgequantum.com
MATTIA FIORENTINI
mattia.fiorentini@cambridgequantum.com
MICHAEL LUBASCH
michael.lubasch@cambridgequanum.com
MATTHIAS ROSENKRANZ
matthias.rosenkranz@cambridgequantum.com

Cambridge Quantum
School of Informatics
University of Edinburgh, United Kingdom
BRIAN COYLE
brian.coyle@cambridgequantum.com

Cambridge Quantum
Terrington House
13-15 Hills Road
Cambridge CB2 1NL
United Kingdom

Variational inference with a quantum computer

Marcello Benedetti,[1, *] Brian Coyle,[1, 2] Mattia Fiorentini,[1] Michael Lubasch,[1] and Matthias Rosenkranz[1, †]

[1]*Cambridge Quantum Computing Limited, SW1E 6DR London, United Kingdom*
[2]*School of Informatics, University of Edinburgh, EH8 9AB Edinburgh, United Kingdom*
(Dated: April 21, 2021)

Inference is the task of drawing conclusions about unobserved variables given observations of related variables. Applications range from identifying diseases from symptoms to classifying economic regimes from price movements. Unfortunately, performing exact inference is intractable in general. One alternative is variational inference, where a candidate probability distribution is optimized to approximate the posterior distribution over unobserved variables. For good approximations a flexible and highly expressive candidate distribution is desirable. In this work, we propose quantum Born machines as variational distributions over discrete variables. We apply the framework of operator variational inference to achieve this goal. In particular, we adopt two specific realizations: one with an adversarial objective and one based on the kernelized Stein discrepancy. We demonstrate the approach numerically using examples of Bayesian networks, and implement an experiment on an IBM quantum computer. Our techniques enable efficient variational inference with distributions beyond those that are efficiently representable on a classical computer.

I. INTRODUCTION

Probabilistic graphical models describe the dependencies of random variables in complex systems [1]. This framework enables two important tasks: learning and inference. Learning yields a model that approximates the observed data distribution. Inference uses the model to answer queries about unobserved variables given observations of other variables. In general, exact inference is intractable and so producing good *approximate* solutions becomes a desirable goal. This article introduces approximate inference solutions using a hybrid quantum-classical framework.

Prominent examples of probabilistic graphical models are Bayesian and Markov networks. Applications across many domains employ inference on those models, including in health care and medicine [2, 3], biology, genetics and forensics [4, 5], finance [6] and fault diagnosis [7]. These are applications where qualifying and quantifying the uncertainty of conclusions is crucial. The posterior distribution can be used to quantify this uncertainty. It can also be used in other downstream tasks such as determining the likeliest configuration of unobserved variables which best explains observed data.

Approximate inference methods broadly fall into two categories: Markov chain Monte Carlo (MCMC) and variational inference (VI). MCMC methods produce samples from the true posterior distribution in an asymptotic limit [8, 9]. VI is a machine learning technique which casts inference as an optimization problem over a parameterized family of probability distributions [10]. If quantum computers can deliver even a small improvement to these methods, the impact across science and engineering could be large.

Initial progress in combining MCMC with quantum computing was achieved by replacing standard MCMC with quantum annealing hardware in the training of some types of Bayesian and Markov networks [11–16]. Despite promising empirical results, it proved difficult to show *if* and *when* a quantum advantage could be delivered. An arguably more promising path towards quantum advantage is through gate-based quantum computers. In this context, there exist algorithms for MCMC with proven asymptotic advantage [17–20], but they require error correction and other features that go beyond the capability of existing machines. Researchers have recently shifted their focus to near-term machines, proposing new quantum algorithms for sampling thermal distributions [21–25].

The advantage of combining quantum computing with VI has not been explored to the same extent of MCMC. Previous work includes classical VI algorithms using ideas from quantum annealing [26–28]. In broad terms, VI exhibits faster convergence than MCMC in high dimensions but it does not provide the same theoretical guarantees.

In this work, we turn our attention to performing VI on a quantum computer. We adopt a distinct approach that focuses on improving inference in classical probabilistic models using quantum resources. We use *Born machines*, which are quantum machine learning models that exhibit high expressivity. We show how to employ gradient-based methods and *amortization* in the training phase. These choices are inspired by recent advances in classical VI [29].

Finding a quantum advantage in machine learning in *any* capacity is an exciting research goal, and promising theoretical works in this direction have been recently developed across several prominent subfields. Advantages in supervised learning have been proposed considering: information theoretic arguments [30–33], probably approximately correct (PAC) learning [34–37] and the representation of data in such models [38–40].

More relevant for our purposes are results in *unsupervised* learning, which have considered complexity and learning theory arguments for distributions [41–43] and quantum non-locality and contextuality [44]. Furthermore, some advantages have been observed experimentally [45–48].
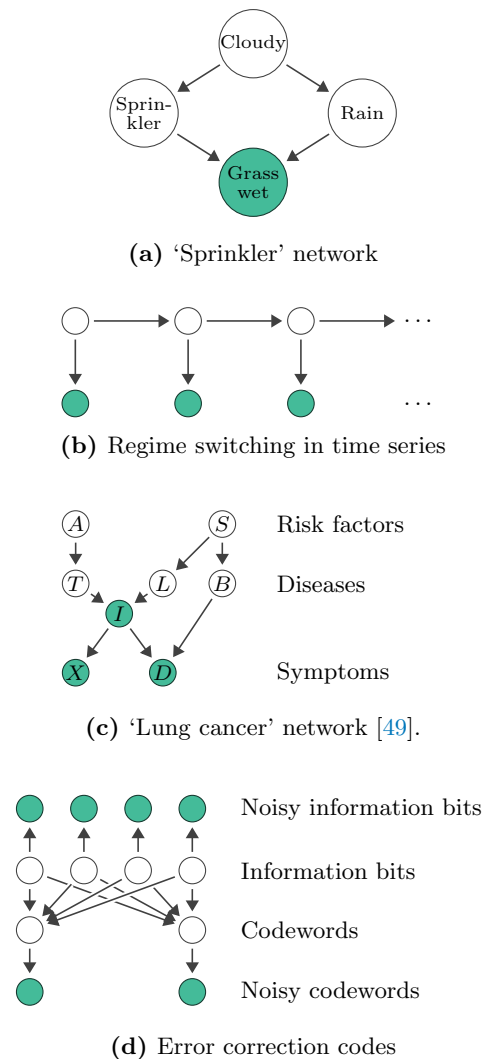
In Section II we describe VI and its applications. In Section III we describe using Born machines to approximate posterior distributions. In Section IV we use the framework of operator VI to derive two suitable objective functions, and we employ classical techniques to deal with the problematic terms in the objectives. In Section V we demonstrate the methods on Bayesian networks. We conclude in Section VI with a discussion of possible generalizations and future work.

II. VARIATIONAL INFERENCE AND APPLICATIONS

It is important to clarify what type of inference we are referring to. Consider a probabilistic model p over some set of random variables, $\mathcal{Y}$. The variables in $\mathcal{Y}$ can be continuous or discrete. Furthermore, assume we are given *evidence* for some variables in the model. This set of variables, denoted $\mathcal{X} \subseteq \mathcal{Y}$, is then *observed* (fixed to the values of the evidence) and we use the vector notation $\boldsymbol{x}$ to denote a realization of these observed variables. We now want to *infer* the posterior distribution of the unobserved variables, those in the set $\mathcal{Z} := \mathcal{Y}\backslash\mathcal{X}$. Denoting these by a vector $\boldsymbol{z}$, our target is that of computing the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$, the conditional probability of $\boldsymbol{z}$ given $\boldsymbol{x}$. By definition, the conditional can be expressed in terms of the joint divided by the marginal: $p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{x},\boldsymbol{z})/p(\boldsymbol{x})$. Also recall that the joint can be written as $p(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. Bayes' theorem combines the two identities and yields $p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})/p(\boldsymbol{x})$.

As described above, the inference problem is rather general. To set the scene, let us discuss some concrete examples in the context of Bayesian networks. These are commonly targeted in inference problems, and as such shall be our test bed for this work. A Bayesian network describes a set of random variables with a clear conditional probability structure. This structure is a directed acyclic graph where the conditional probabilities are modeled by tables, by explicit distributions, or even by neural networks. Figure 1a is the textbook example of a Bayesian network for the distribution of binary variables: cloudy (C), sprinkler (S), rain (R) and the grass being wet (W). According to the graph this distribution factorizes as $P(C,S,R,W) = P(C)P(S|C)P(R|C)P(W|S,R)$. A possible inferential question is: what is the probability distribution of C, S and R given that $W = \mathsf{tr}$? This can be estimated by 'inverting' the probabilities using Bayes' theorem: $p(C,S,R|W = \mathsf{tr}) = p(W = \mathsf{tr}\,|C,S,R)p(C,S,R)/p(W)$.

Figures 1b–1d show a few additional applications of inference on Bayesian networks. The hidden

(a) 'Sprinkler' network

(b) Regime switching in time series

(c) 'Lung cancer' network [49].

(d) Error correction codes

Figure 1: Some applications of inference on Bayesian networks. Filled (empty) circles indicate observed (unobserved) random variables.

Markov model in Fig. 1b describes the joint probability distribution of a time series of asset returns and an unobserved 'market regime' (e.g. a booming vs a recessive economic regime). A typical inference task is to detect regime switches by observing asset returns [50]. Figure 1c illustrates a modified version of the 'lung cancer' Bayesian network which is an example from medical diagnosis (see e.g. [3] and references therein). This network encodes expert knowledge about the relationship between risk factors, diseases and symptoms. In health care, careful design of the network and algorithm are critical in order to reduce biases, e.g. in relation to health care access [51]. Note that inference in medical diagnosis is often causal instead of associative [3]. Bayesian networks can be interpreted causally and help answer causal queries using Pearl's do-calculus [52]. Finally, Figure 1d shows a

Bayesian network representation of turbo codes, which are error correction scheme used in 3G and 4G mobile communications. The inference task for the receiver is to recover the original information bits from the information bits and codewords received over a noisy channel [53].

Inference is a computationally hard task in all but the simplest probabilistic models. Roth [54] extended results of Cooper [55] and showed that *exact* inference in Bayesian networks with discrete variables is #P-complete. Dagum and Luby [56] showed that even *approximate* inference is NP-hard. Therefore, unless some particular constraints are in place, these calculations are intractable. In many cases one is able to perform a 'forward pass' and obtain unbiased samples from the joint $(\boldsymbol{x}, \boldsymbol{z}) \sim p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. However, obtaining unbiased samples from the posterior $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{z})/p(\boldsymbol{x})$ is intractable due to the unknown normalization constant. One can perform MCMC sampling by constructing an ergodic Markov chain whose stationary distribution is the desired posterior. MCMC methods have nice theoretical guarantees, but they may converge slowly in practice [9]. In contrast, VI is often faster in high dimensions but does not come with guarantees. The idea in VI is to optimize a variational distribution q by minimizing its 'distance' from the true posterior p (see [10] for an introduction to the topic and [29] for a review of the recent advances).

VI has experienced a resurgence in recent years due to substantial developments. First, generic methods have reduced the amount of analytical calculations required and have made VI much more user friendly (e.g. black-box VI [57]). Second, machine learning has enabled the use of highly expressive variational distributions implemented via neural networks [58], probabilistic programs [59], non-parametric models [60], normalizing flows [61] and others. In contrast, the original VI methods were mostly limited to analytically tractable distributions such as those which could be factorized. Third, amortization methods have reduced the costs by optimizing $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ where the vector of parameters $\boldsymbol{\theta}$ is 'shared' among all possible observations $\boldsymbol{x}$ instead of optimizing individual parameters for each observation. This approach can also *generalize* across inferences.

Casting the inference problem as an optimization problem comes itself with challenges, in particular when the unobserved variables are discrete. REINFORCE [62] is a generic method which requires calculation of the score $\partial_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ and may suffer from high variance. Gumbel-softmax reparametrizations [63, 64] use a continuous relaxation which does not follow the exact distribution.

We now show that near-term quantum computers provide an alternative tool for VI. We use the quantum Born machine as a candidate for variational distributions. These models are highly expressive and they naturally represent discrete distributions as a result of quantum measurement. Furthermore, the Born machine can be trained by gradient-based optimization.
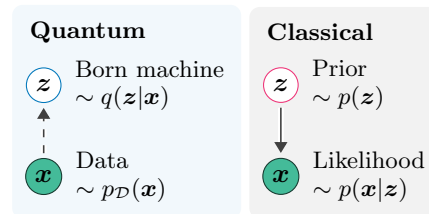
Figure 2: Probabilistic models used in our methods. The classical model comprises a prior over unobserved discrete variables and a likelihood over observed variables. The quantum model approximates the posterior distribution of the unobserved variables given observed data. All distributions can be sampled efficiently as long as one follows the arrows and uses a suitable computer.

III. BORN MACHINES AS IMPLICIT VARIATIONAL DISTRIBUTIONS

By exploiting the inherent probabilistic nature of quantum mechanics, one can model the probability distribution of classical data using a pure quantum state. This model based on the Born rule in quantum mechanics is referred to as Born machine [65]. Let us consider binary vectors $\boldsymbol{z} \in \{0, 1\}^n$ where n is the number of variables. The Born machine is a normalized quantum state $|\psi(\boldsymbol{\theta})\rangle$ parameterized by $\boldsymbol{\theta}$ which outputs n-bit strings with probabilities $q_{\boldsymbol{\theta}}(\boldsymbol{z}) = |\langle \boldsymbol{z}|\psi(\boldsymbol{\theta})\rangle|^2$. Here $|\boldsymbol{z}\rangle$ are computational basis states, thus sampling the above probability boils down to a simple measurement. Other forms of discrete variables can be dealt with by a suitable encoding. When using amortization, the variational distribution requires conditioning on observed variables. To extend the Born machine and include this feature, we let $\boldsymbol{x}$ play the role of additional parameters. This yields a pure state where the output probabilities are $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = |\langle \boldsymbol{z}|\psi(\boldsymbol{\theta}, \boldsymbol{x})\rangle|^2$. Figure 2 shows the relation between the classical model and the quantum model for approximate inference.

Born machines have been applied for benchmarking hybrid quantum-classical systems [66–69], generative modeling [70–72], finance [46, 47, 73], anomaly detection [74], and have been proposed for demonstrating quantum advantage [42][1]. These models can be realized in a variety of ways, in both classical and quantum computers. When realized via certain classes of quantum circuits they are classically intractable to simulate.[2]

---

[1] Some literature (e.g. [73]) refer to Born machines as quantum generative adversarial networks (QGANs) when using an adversarial objective function. It is more appropriate to reserve the term QGAN to adversarial methods for quantum data (e.g. [75–77]).

[2] A quantum computer cannot efficiently implement arbitrary circuits, thus some Born machines are intractable even for a quantum computer.

For example, instantaneous quantum poly-time (IQP) circuits are Born machines with $O(\text{poly}(n))$ parameters that yield classically intractable distributions in the average case under widely accepted complexity theoretic assumptions [78]. Additional examples for such classically hard circuits are boson sampling [79] and random circuits [80, 81]. Thus quantum Born machines have expressive power larger than that of classical models, including neural networks [82] and partially matrix product states [83]. It can be shown that the model remains classically intractable throughout training [42].

A useful way to classify probabilistic models is the following: *prescribed* models provide an explicit parametric specification of the distribution, *implicit* models define only the data generation process [84]. Born machines can be effectively regarded as implicit models. It is easy to obtain an unbiased sample as long as we can execute the corresponding circuit on a quantum computer and measure the computational basis. However, it requires exponential resources to estimate the distribution even on a quantum computer. To see this consider the conditional distribution $q_{\boldsymbol{\theta}}(\boldsymbol{z}_a|\boldsymbol{z}_b, \boldsymbol{x})$ where $(\boldsymbol{z}_a, \boldsymbol{z}_b)$ is a partition of $\boldsymbol{z}$. To estimate this from samples of the joint distribution we would need the unnatural resource of post-selection [85]. Simulating post-selection has exponential cost in the worst case. There exist classes of Born machines for which the dependency between parameters and distribution is clear. For example, when restricting the Born machine to an IQP circuit [42], amplitudes $\langle \boldsymbol{z}|\psi(\boldsymbol{\theta})\rangle$ are proportional to partition functions of complex Ising models [86, 87]. Yet estimation of arbitrary amplitudes remains intractable.

Implicit models are challenging to train with standard methods. The major challenge is the design of the objective function precisely because likelihoods are 'prohibited'. Valid objectives involve only statistical quantities (such as expectation values) that can be efficiently estimated from samples.[3] For generative modeling with Born machines progress has been made towards practical objectives such as moment-matching [66], maximum mean discrepancy [88], Stein and Sinkhorn divergences [42], adversarial objectives [89, 90] as well as incorporating Bayesian priors on the parameters [82].

In the next section we make some progress towards practical objectives for VI with Born machines. First, we mention some related work. The approaches in [22] use VI ideas to deal with thermal states and quantum data. In contrast, our inference methods apply to classical graphical models and classical data. The approaches of [17, 91, 92] aim to encode Bayesian networks directly in a quantum state, and subsequently perform exact inference. In contrast, our inference methods are

approximate, efficient, and apply also to graphical models that are not Bayesian networks.

IV. OPERATOR VARIATIONAL INFERENCE

Operator Variational Inference (OPVI) [93] is a rather general method that uses mathematical operators to design objectives for the approximate posterior. Suitable operators are those for which (i) the minima of the variational objective is attained at the true posterior, and (ii) it is possible to estimate the objective without computing the true posterior. In general, the amortized OPVI objective is:

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}(\boldsymbol{x})} \sup_{f\in\mathcal{F}} h\left(\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[(O^{p,q}f)(\boldsymbol{z})\right]\right), \quad (1)$$

where $f(\cdot) \in \mathbb{R}^d$ is a test function within the chosen family $\mathcal{F}$, $O^{p,q}$ is an operator that depends on $p(\boldsymbol{z}|\boldsymbol{x})$ and $q(\boldsymbol{z}|\boldsymbol{x})$, and $h(\cdot) \in [0,\infty]$ yields a non-negative objective. Note that we have an expectation over the data distribution $p_{\mathcal{D}}(\boldsymbol{x})$. This indicates the average over a dataset $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_i$ of observations to condition on.

We present two methods that follow directly from two operator choices. These operator choices result in objectives based on the Kullback-Leibler (KL) divergence and the Stein discrepancy. The former utilizes an *adversary* to make the computation tractable, whereas in the latter, the tractability arises from the computation of a *kernel* function.

The KL divergence is an example of an *f-divergence*, while the Stein discrepancy is in the class of *integral probability metrics*, two fundamentally different families of probability distance measures. OPVI can therefore yield methods from these two different classes under a suitable choice of operator. However, it is shown in [94] that these two families intersect non-trivially only at the total variation distance (TVD). It is for this reason that we choose the TVD as our benchmark in later numerical results.

A. The adversarial method

One possible objective function for VI is the Kullback-Leibler divergence (KL) of the true posterior relative to the approximate one. This is obtained from Eq. (1) by choosing f and h to be identity functions, and by choosing the operator $(O^{p,q}f)(\boldsymbol{z}) = \log\frac{q(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})}$:

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}(\boldsymbol{x})} \underbrace{\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})}\right]}_{\text{KL}[q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}|\boldsymbol{x})]}. \quad (2)$$

Here $q_{\boldsymbol{\theta}}$ is the variational distribution parameterized by $\boldsymbol{\theta}$. The objective's minimum is zero and is attained by the true posterior $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}), \forall \boldsymbol{x}$. There are

---

[3] Bias and variance of the estimates need to be analyzed and controlled, which is a challenge in itself.
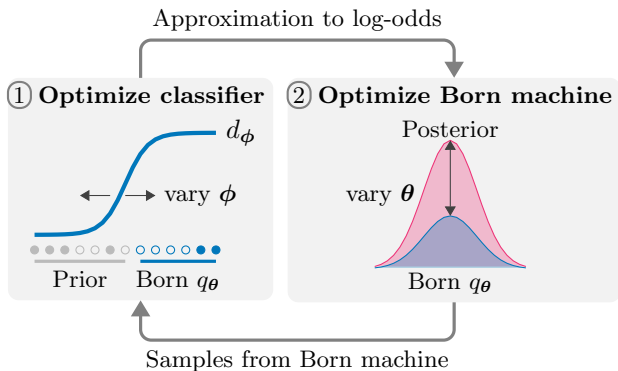
Figure 3: Adversarial variational inference with a Born machine. Step 1 optimizes a classifier d_ϕ to output the probabilities that the observed samples come from the Born machine rather than from the prior. Step 2 optimizes the Born machine q_θ to better match the true posterior. The updated Born machine is fed back into step 1 and the process repeats until convergence of the Born machine.

many ways to rewrite this objective. Here we focus on the form known as *prior-contrastive*. Substituting Bayes formula $p(z|x) = p(x|z)p(z)/p(x)$ in the equation above we obtain:

$$\mathbb{E}_{x \sim p_\mathcal{D}(x)}\mathbb{E}_{z \sim q_\theta(z|x)}\left[\log \frac{q_\theta(z|x)}{p(z)} - \log p(x|z)\right] + \text{const.}, \tag{3}$$

where we ignore the term $-\mathbb{E}_{x \sim p_\mathcal{D}(x)}[\log p(x)]$ since it is constant with respect to q. This objective has been used in many generative models, including the celebrated variational autoencoder (VAE) [58, 95]. While the original VAE relies on tractable variational posteriors, one could also use more powerful, implicit distributions as demonstrated in adversarial variational Bayes [96] and in prior-contrastive adversarial VI [97].

Let us use a Born machine to model the variational distribution $q_\theta(z|x) = |\langle z|\psi(\theta, x)\rangle|^2$. As the model is implicit, the ratio $q_\theta(z|x)/p(z)$ in Eq. (3) cannot be computed efficiently. Therefore, we introduce an adversarial method for estimating the ratio above approximately. The key insight here is that this ratio can be estimated from the output of a binary classifier [84]. Suppose we ascribe samples $(z, x) \sim q_\theta(z|x)p_\mathcal{D}(x)$ to the first class, and samples $(z, x) \sim p(z)p_\mathcal{D}(x)$ to the second class. A binary classifier d_ϕ parameterized by ϕ outputs the probability $d_\phi(z, x)$ that the pair (z, x) belongs to one of two classes. Hence, $1 - d_\phi(z, x)$ indicates the probability that (z, x) belongs to the other class. There exist many possible choices of objective function for the classifier [84]. In this work we consider the cross-entropy:

$$\begin{aligned}\mathcal{G}_{\mathrm{KL}}(\phi; \theta) = &\mathbb{E}_{x \sim p_\mathcal{D}(x)}\mathbb{E}_{z \sim q_\theta(z|x)}\left[\log d_\phi(z, x)\right] \\ &+ \mathbb{E}_{x \sim p_\mathcal{D}(x)}\mathbb{E}_{z \sim p(z)}\left[\log(1 - d_\phi(z, x))\right].\end{aligned} \tag{4}$$

The optimal classifier which maximizes this equation is [96]:

$$d^*(z, x) = \frac{q_\theta(z|x)}{q_\theta(z|x) + p(z)}. \tag{5}$$

Since the probabilities in Eq. 5 are unknown, the classifier must be trained on a dataset of samples. This does not pose a problem because samples from the Born machine $q_\theta(z|x)$ and prior $p(z)$ are easy to obtain by assumption. Once the classifier is trained, the logit transformation provides the log-odds of a data point coming from the Born machine joint $q_\theta(z|x)p_\mathcal{D}(x)$ vs the prior joint $p(z)p_\mathcal{D}(x)$. The log-odds are an approximation to the log-ratio of the two distributions:

$$\mathrm{logit}(d_\phi(z, x)) \equiv \log \frac{d_\phi(z, x)}{1 - d_\phi(z, x)} \approx \log \frac{q_\theta(z|x)}{p(z)}, \tag{6}$$

which is exact if d_ϕ is the optimal classifier in Eq. (5). Now, we can avoid the computation of the problematic term in the KL divergence. Plugging this result in Eq. (3) and ignoring the constant term, the final objective for the Born machine is:

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}}(\theta; \phi) = &\mathbb{E}_{x \sim p_\mathcal{D}(x)}\mathbb{E}_{z \sim q_\theta(z|x)}\left[\mathrm{logit}(d_\phi(z, x))\right. \\ &\left. - \log p(x|z)\right].\end{aligned} \tag{7}$$

The optimization can be performed in tandem as:

$$\begin{aligned}&\max_\phi \mathcal{G}_{\mathrm{KL}}(\phi; \theta) \\ &\min_\theta \mathcal{L}_{\mathrm{KL}}(\theta; \phi),\end{aligned} \tag{8}$$

using gradient ascent and descent, respectively. It can be shown [96, 97] that the gradient of $\log \frac{q_\theta(z|x)}{p(z)}$ with respect to θ vanishes. Thus, under the assumption of an optimal classifier d_ϕ, the gradient of Eq. (7) is significantly simplified. This gradient is derived in Appendix A.

A more intuitive interpretation of the procedure just described is as follows. The log-likelihood in Eq. (3) can be expanded as $\log p(x|z) = \log \frac{p(z|x)}{p(z)} + \log p(x)$. Then Eq. (3) can be rewritten as $\mathbb{E}_{x \sim p_\mathcal{D}(x)}\mathbb{E}_{z \sim q_\theta(z|x)}[\log \frac{q_\theta(z|x)}{p(z)} - \log \frac{p(z|x)}{p(z)}]$. Comparing the expression in the square brackets with Eq. (6) reveals that the difference is between two log-odds: the first is given by an optimal classifier for the approximate posterior and prior, and the second is given by a hypothetical classifier for the true posterior and prior. The adversarial method is illustrated in Fig. 3.

B. The kernelized method

Another possible objective function for VI is the Stein discrepancy (SD) of the true posterior from the approximate one. This is obtained from Eq. (1) assuming that the image of f has the same dimension as z, choosing
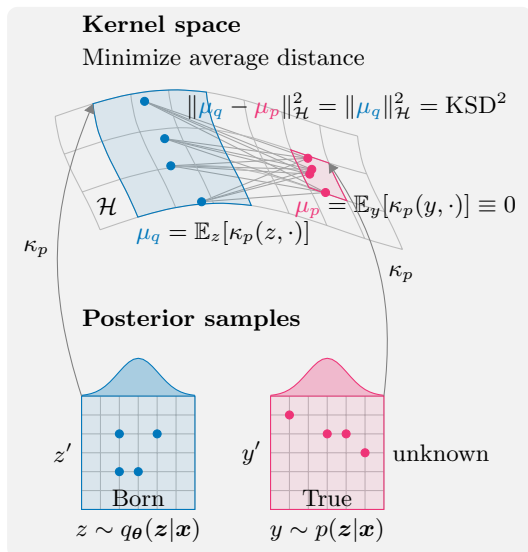
Figure 4: Kernelized Stein variational inference with a Born machine. This method minimizes the kernelized Stein discrepancy between the approximate and true posterior distributions. First, both distributions are embedded in a functional Hilbert space $\mathcal{H}$. The map κ_p is designed so that elements mapped from the unknown true posterior evaluate to zero. Second, the discrepancy is evaluated using samples from the Born machine $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ and optimized by gradient descent.

h to be the absolute value, and choosing $O^{p,q}$ to be a Stein operator. A Stein operator is independent from q and is characterized by having zero expectation under the true posterior for all functions f in the chosen family $\mathcal{F}$. Then, for binary variables a possible Stein operator is $(O^p f)(\boldsymbol{z}) = s_p(\boldsymbol{x}, \boldsymbol{z})^T f(\boldsymbol{z}) - \text{tr}(\Delta f(\boldsymbol{z}))$ where:

$$(s_p(\boldsymbol{x}, \boldsymbol{z}))_i = 1 - \frac{p(\boldsymbol{x}, \neg_i \boldsymbol{z})}{p(\boldsymbol{x}, \boldsymbol{z})}, \tag{9}$$

is the difference score function,

$$(\Delta f(\boldsymbol{z}))_{ij} = (f(\boldsymbol{z}))_j - (f(\neg_i \boldsymbol{z}))_j, \tag{10}$$

is the partial difference operator, and $\neg_i \boldsymbol{z}$ flips the i-th bit in binary vector $\boldsymbol{z}$. We show in Appendix B that this is a valid Stein operator for binary variables. For more general discrete variables we refer the interested reader to [98]. Plugging these definitions in Eq. (1) we obtain:[4]

$$\mathbb{E}_{\boldsymbol{x} \sim p_{\mathcal{D}}(\boldsymbol{x})} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})} \left[s_p(\boldsymbol{x}, \boldsymbol{z})^T f(\boldsymbol{z}) - \text{tr}(\Delta f(\boldsymbol{z})) \right] \right|.$$

$$\underbrace{}_{\text{SD}[q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}|\boldsymbol{x})]}$$

$$\tag{11}$$

––––––––––

[4] In contrast to [98] but in line with earlier work such as [99] we include the absolute value in the definition of the Stein discrepancy.

At this point one can parameterize the test function f and obtain an adversarial objective similar in spirit to the one presented in Section IV A. Here, however, we take a different route. By restricting the Hilbert space norm of f to be at most 1, the supremum in Eq. (11) can be calculated in closed form using a kernel [98]. The result is the kernelized Stein discrepancy (KSD):

$$\text{KSD}\left[q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}|\boldsymbol{x})\right] = \sqrt{\mathbb{E}_{\boldsymbol{z}, \boldsymbol{z}' \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}[\kappa_p(\boldsymbol{z}, \boldsymbol{z}')]}, \tag{12}$$

where κ_p is a Stein kernel. For binary variables this can be written as:

$$\kappa_p(\boldsymbol{z}, \boldsymbol{z}') = s_p(\boldsymbol{x}, \boldsymbol{z})^T k(\boldsymbol{z}, \boldsymbol{z}') s_p(\boldsymbol{x}', \boldsymbol{z}')$$
$$- s_p(\boldsymbol{x}, \boldsymbol{z})^T \Delta_{\boldsymbol{z}'} k(\boldsymbol{z}, \boldsymbol{z}') - \Delta_{\boldsymbol{z}} k(\boldsymbol{z}, \boldsymbol{z}')^T s_p(\boldsymbol{x}', \boldsymbol{z}')$$
$$+ \text{tr}(\Delta_{\boldsymbol{z}, \boldsymbol{z}'} k(\boldsymbol{z}, \boldsymbol{z}')). \tag{13}$$

The Stein kernel, κ_p, depends on *another* kernel k. For n unobserved Bernoulli variables one possibility is the generic Hamming kernel $k(\boldsymbol{z}, \boldsymbol{z}') = \exp\left(-\frac{1}{n}\|\boldsymbol{z} - \boldsymbol{z}'\|_1\right)$. The KSD is a valid discrepancy measure if this 'internal' kernel, k, is positive definite which is the case for the Hamming kernel [98].

In summary, constraining $\|f\|_{\mathcal{H}} \leq 1$ in Eq. (11) and substituting the KSD we obtain:

$$\mathcal{L}_{\text{KSD}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x} \sim p_{\mathcal{D}}(\boldsymbol{x})} \sqrt{\mathbb{E}_{\boldsymbol{z}, \boldsymbol{z}' \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}[\kappa_p(\boldsymbol{z}, \boldsymbol{z}')]}, \tag{14}$$

and the problem consists of finding $\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{KSD}}(\boldsymbol{\theta})$. The gradient of Eq. (14) is derived in Appendix C. The kernelized method is illustrated in Fig. 4.

The KSD was used in [42] for generative modeling with Born machines. In that context the distribution p is unknown, thus the authors derived methods to approximate the score s_p from available data. VI is a suitable application of the KSD because in this context we do know the joint $p(\boldsymbol{z}, \boldsymbol{x})$. Moreover the score in Eq. (9) can be computed efficiently even if we have the joint in unnormalized form.

V. EXPERIMENTS

To demonstrate our approach, we employ three experiments as the subject of the following sections. First, we validate both methods using the canonical 'sprinkler' Bayesian network. Second, we extend the regime of applications by considering *continuous* observed variables in a hidden Markov model (HMM), and illustrate how multiple observations can be incorporated effectively via amortization. Finally, we consider a larger discrete model in the 'lung cancer' Bayesian network and demonstrate the methods using an IBM quantum computer.
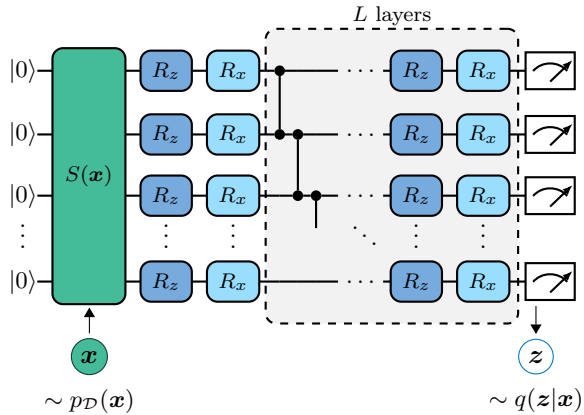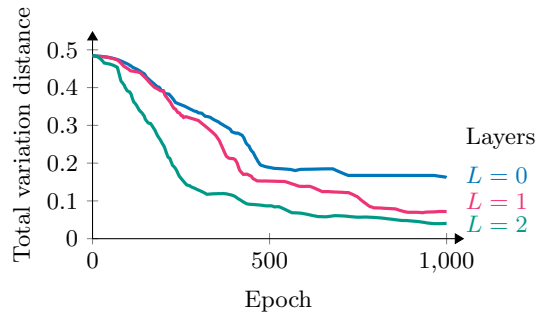
Figure 5: Hardware-efficient ansatz for the Born machine used in the experiments. All rotations R_x, R_z are parameterized by individual angles. The unitary $S(\boldsymbol{x})$ encodes the values of $\boldsymbol{x}$. The particular choice of $S(\boldsymbol{x})$ depends on the application (see the main text). The corresponding classical random variables are indicated below the circuit.

(a) Adversarial objective

(b) Kernelized Stein objective

Figure 6: Total variation distance of the approximate and true posteriors. Median of 30 instances for (a) KL and (b) KSD objective functions for a different number of layers in the ansatz (cf. Fig. 5). Each instance is a 'sprinkler' Bayesian network as in Fig. 1a where the conditional probabilities are chosen at random.
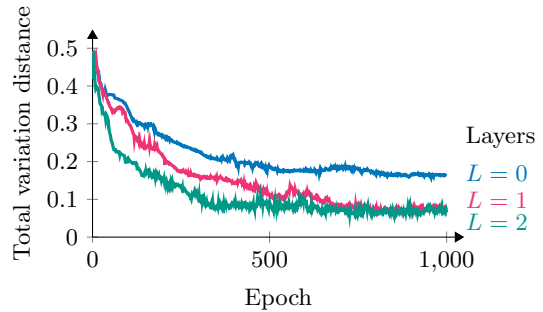
A.   Proof of principle with the 'sprinkler' Bayesian network

As a first experiment we classically simulate the methods on the 'sprinkler' network in Fig. 1a, one of the simplest possible examples. The purpose is to show that both methods are expected to work well even when using limited resources and without much fine-tuning. First, we randomly generate the entries of each probability table from the uniform distribution $\mathcal{U}([0.01, 0.99])$ and produced a total of 30 instances of this network. For each instance we condition on 'Grass wet' being true which means our data distribution is $p_{\mathcal{D}}(W) = \delta(W = \mathsf{tr})$. We infer the posterior of the remaining variables using a Born machine on 3 qubits and with the hardware-efficient ansatz shown in Fig. 5. We use a layer of Hadamard gates as a state preparation, $S(\boldsymbol{x}) = H \otimes H \otimes H$, and initialized all parameters to ≈ 0. Thus the initial distribution is approximately uniform over all bit-strings and does not make any assumption. Such hardware-efficient ansätze, while simple, have been shown to be vulnerable to *barren plateaus* [100–103], or regions of exponentially vanishing gradient magnitudes which makes training untenable. Alternatively, one could consider ansätze which have been shown to be somewhat 'immune' to this phenomenon [104, 105], but we leave such investigation to future work.

For the KL objective, we utilize a multi-layer perceptron (MLP) classifier made of 3 input units, 6 ReLU hidden units and 1 sigmoid output unit. The classifier is trained with a dataset of 100 samples from the prior $p(C, R, S)$ and 100 samples from the Born machine $q(C, R, S|W = \mathsf{tr})$. Here we use stochastic gradient descent with batches of size 10 and a learning rate of

0.03. For the KSD objective, we use the Hamming kernel as above. For both cases, we compute the total variation distance (TVD) of true and approximate posterior at each epoch. We emphasize that TVD cannot be efficiently computed in general and can be shown only for small examples. Figure 6a (Figure 6b) shows the median TVD out of the 30 instances for 1000 epochs of training for the KL (KSD) objective respectively. For both methods, the Born machine is trained using 100 samples to estimate each expectation value for the gradients, and using vanilla gradient descent with a small learning rate of 0.003. The 0-layers Born machine generates unentangled states and it can be thought of as a classical mean-field approximation. Increasing number of layers lead to better approximations to the posterior in all cases.

Qualitatively, KL tends to converge to slightly better approximate posteriors, but requires more memory and computation than KSD. This is because of the MLP classifier which is trained along the Born machine. In Appendix D we provide additional results to supplement those in Fig. 6. We do not attempt to compare the two methods quantitatively as it would require the fine-tuning of a large number of hyper-parameters and we leave this study for future work.

B. Continuous observed variables and amortization with a hidden Markov model

In our second experiment, we simulate the adversarial VI method on the hidden Markov model (HMM) in Fig. 1b. The purpose was to demonstrate two interesting features: continuous observed variables and amortization. We set up the HMM for $T = 8$ time steps (white circles in Fig. 1b), each represented by a Bernoulli latent variable with conditional dependency:

$$z_1 \sim \mathcal{B}(\tfrac{1}{2}),$$
$$z_t \sim \begin{cases} \mathcal{B}(\tfrac{1}{3}) & \text{if } z_{t-1} = 0 \\ \mathcal{B}(\tfrac{2}{3}) & \text{if } z_{t-1} = 1. \end{cases} \qquad (15)$$

These represent the unknown 'regime' at time t. The regime affects how the observable data is generated. We use Gaussian observed variables (green circles in Fig. 1b) whose mean and standard deviation depends on the latent variables as:

$$x_t \sim \begin{cases} \mathcal{N}(0, 1) & \text{if } z_t = 0 \\ \mathcal{N}(1, \tfrac{1}{2}) & \text{if } z_t = 1. \end{cases} \qquad (16)$$

We sample two time series observations $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$ from the HMM and take this to be our data distribution. These time series are shown by green dots in Fig. 7. Instead of fitting two approximate posteriors *separately*, we use a single Born machine with amortization $|\psi(\boldsymbol{\theta}, \boldsymbol{x})\rangle$. We used the ansatz in Fig. 5 for 8 qubits with a state preparation layer $S(\boldsymbol{x}) = \otimes_{t=1}^{8} R_x(x_t)$. In practice, this encoding step can be done under more careful considerations [106, 107]. Parameters $\boldsymbol{\theta}$ are initialized to small values at random. We optimize the KL objective and a MLP classifier with 16 inputs, 24 ReLU, and 1 sigmoid units and the system is trained for 3000 epochs. The learning rates are set to be 0.006 for the Born machine and 0.03 for the MLP respectively. The Born machine used 100 samples to estimate each expectation value, the MLP used 100 samples from each distribution and mini-batches of size 10.

The histograms in Fig. 7 show the 10 most probable configurations of latent variables for the true posterior along with the probabilities assigned by the Born machine. Conditioning on datapoint $\boldsymbol{x}^{(1)}$, the inferred most likely explanation is $|01100011\rangle$ (i.e. the mode of the Born machine). This coincides with the true posterior mode. For $\boldsymbol{x}^{(2)}$, the inferred mode was $|10001000\rangle$, which differs from the true posterior mode $|10000000\rangle$ by a single bit. The regime switching has therefore been modeled with high accuracy.

Rather than focusing on the mode, one can make use of the whole distribution to estimate some quantity of interest. This is done by taking samples from the Born machine and using them in a Monte Carlo estimate for such a quantity. For example, we could predict the expected value of the next latent variable z_{T+1} given the

available observations. For datapoint $\boldsymbol{x}^{(1)}$ this entails the estimation of:

$$\mathbb{E}_{z_T \sim q_\theta(z_T | \boldsymbol{x}^{(1)})} \mathbb{E}_{z_{T+1} \sim p(z_{T+1} | z_T)} [z_{T+1}]. \qquad (17)$$

To conclude, we mention that inference in this HMM can be performed exactly and efficiently with classical algorithms. The most likely explanation for the observed data can be found with the Viterbi algorithm in time $O(T|S|^2)$ where T is the length of the sequence and $|S|$ is the size of the unobserved variables (e.g. $|S| = 2$ for Bernoulli variables). This is not the case for more general models. For example, a factorial HMM has multiple independent chains of unobserved variables [108]. Exact inference costs $O(TM|S|^{M+1})$ where M is number of chains, and is typically replaced by approximate inference [108, 109]. Our VI methods are generic and apply to factorial and other HMMs without changes.

C. Demonstration on IBMQ with the 'lung cancer' Bayesian network

As a final experiment, we test the performance on real quantum hardware using the slightly more complex 'lung cancer' network [49], specifically using the 5 qubit `ibmq_rome` quantum processor. We access this device using `PyQuil` [110] and `tket` [111].
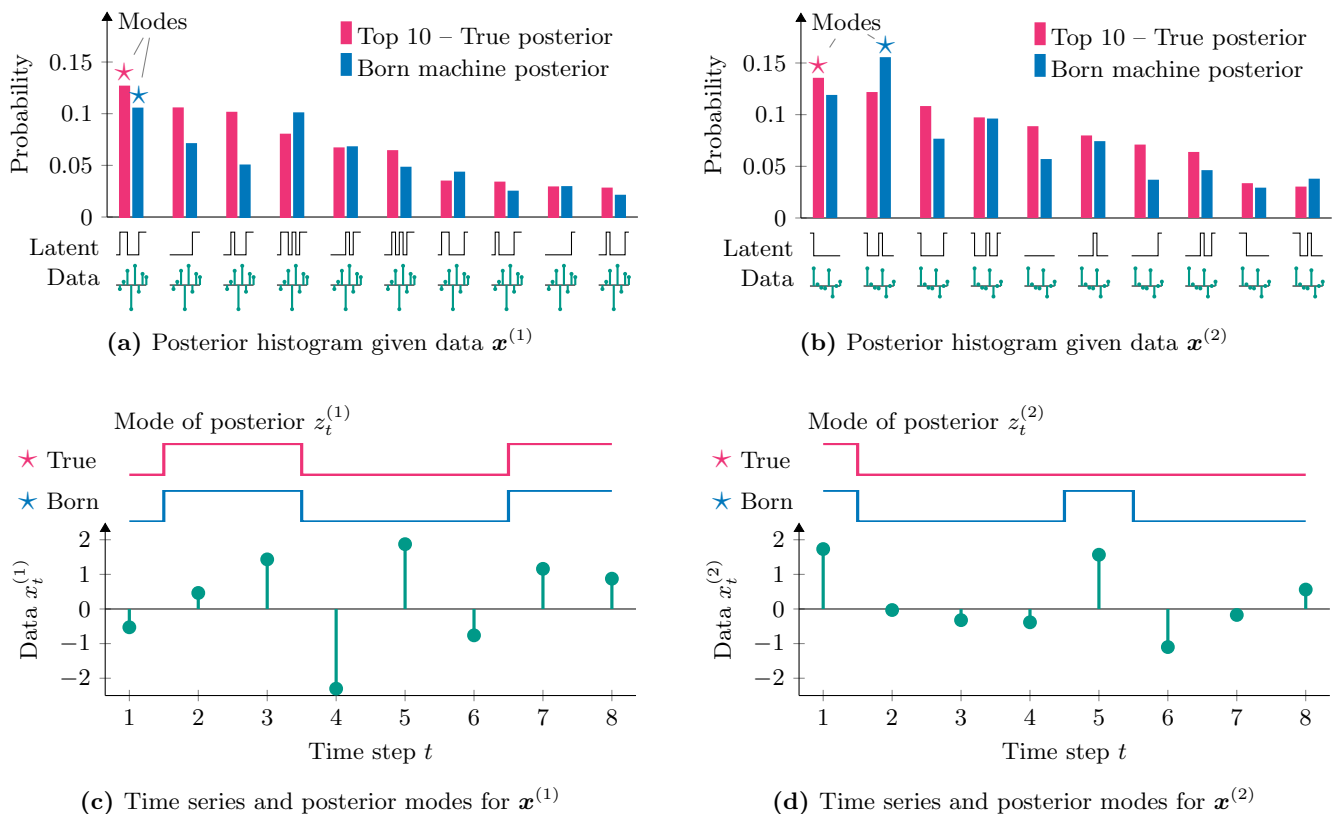
The lung cancer network (also called the 'Asia' network) is an example of a medical diagnosis Bayesian network, and is illustrated by Fig. 1c. This network was chosen since it is small enough to fit without adaptation onto the quantum device, but also comes closer to a 'real-world' example. A more complicated version with extra variables can be found in [112]. We note that the version we use is slightly modified from that of [49] for numerical reasons.

The network has 8 possible nodes: two 'symptoms', whether the patient presented with dyspnoea (D) (shortness of breath), or had a positive X-ray (X); 4 possible 'diseases' causing the symptoms: bronchitis (B), tuberculosis (T), lung cancer (L), or an 'illness' (I) (which could be either tuberculosis or lung cancer or something other than bronchitis); and two possible 'risk factors': whether the patient had traveled to Asia (A), or whether they had a history of smoking (S).

Based on the graph structure in Fig. 1c, the distribution over the variables, $p(A, T, S, L, I, X, B, D)$ can be factorized as:

$$p(A)p(T|A)p(X|I)p(I|T,L)p(D|B,I)p(B|S)p(L|S)p(S) \qquad (18)$$

In Appendix E we show the explicit probability table we use (Table I) for completeness. Modifying an illustrative example of a potential 'real-world' use case in [49], a patient may present in a clinic with an 'illness' ($I = \mathsf{tr}$) but no shortness of breath ($D = \mathsf{fa}$) as a symptom. Furthermore, an x-ray has revealed a negative result

(a) Posterior histogram given data $\boldsymbol{x}^{(1)}$

(b) Posterior histogram given data $\boldsymbol{x}^{(2)}$

(c) Time series and posterior modes for $\boldsymbol{x}^{(1)}$

(d) Time series and posterior modes for $\boldsymbol{x}^{(2)}$

Figure 7: Truncated, ordered histograms of the posteriors for two observed samples (a) $\boldsymbol{x}^{(1)}$ and (b) $\boldsymbol{x}^{(2)}$ of the hidden Markov model in Eq. (15)–(16). The histograms are sorted by probability of the true posterior. The blue bars are the probabilities of the corresponding approximate posterior. The x-axis shows the latent state for each bar and the observed data point $\boldsymbol{x}$. The lower panels are the time series of the data (c) $\boldsymbol{x}^{(1)}$ and (d) $\boldsymbol{x}^{(2)}$, as well as the corresponding modes of the true posterior and Born machine posterior as indicated with stars in the upper panel.

($X = \mathsf{fa}$). However, we have no patient history and we do not know which underlying disease is actually present. As such in the experiment, we condition on having observed the 'evidence' variables, $\boldsymbol{x}$: X, D and I. The remaining five are the latent variables, $\boldsymbol{z}$, so we will require 5 qubits.

In Fig. 8, we show the results when using the 5 qubit `ibmq_rome` quantum processor, which has a linear topology. The topology is convenient since it introduces no major overheads when compiling from our ansatz in Fig. 5. We plot the true posterior versus the one learned by the Born machine both simulated, and on the quantum processor using the best parameters found (after ~ 400 epochs simulated and ~ 50 epochs on the processor). To train in both cases, we use the same classifier as in Section V B with 5 inputs and 10 ReLU hidden units using the KL objective, which we observed to give the best results. We employ a two layer ansatz in Fig. 5 ($L = 2$), and 1024 shots are taken from the Born machine in all cases. We observe that the simulated Born machine is able to learn the true posterior very well with these parameters, but the performance suffers on the hardware. That being said, the trained hardware model is still able to pick three out of four highest probability configurations of the network (shown in the x-axis of

Fig. 8). The hardware results could likely be improved with error mitigation techniques.

VI. DISCUSSION

We presented two variational inference (VI) methods that can exploit highly expressive approximate posteriors given by quantum models. The first method is based on minimizing the Kullback-Leibler (KL) divergence to the true posterior and relies on a classifier that estimates probability ratios. The resulting adversarial training may be challenging due to the large number of hyper-parameters and well-know stability issues [113, 114]. On the other hand, it only requires the ability to (i) sample from the prior $p(\boldsymbol{z})$, (ii) sample from the Born machine $q_\theta(\boldsymbol{z}|\boldsymbol{x})$, and (iii) calculate the likelihood $p(\boldsymbol{x}|\boldsymbol{z})$. We can apply the method as-is even when the prior is implicit (e.g. in the quantum-assisted Helmholtz machine [15]). If the likelihood is also implicit we can reformulate Eq (3) in *joint contrastive* form [97] and apply the method with minor changes. This opens the possibility to train models where both the generative and inference processes are described by Born machines (i.e.
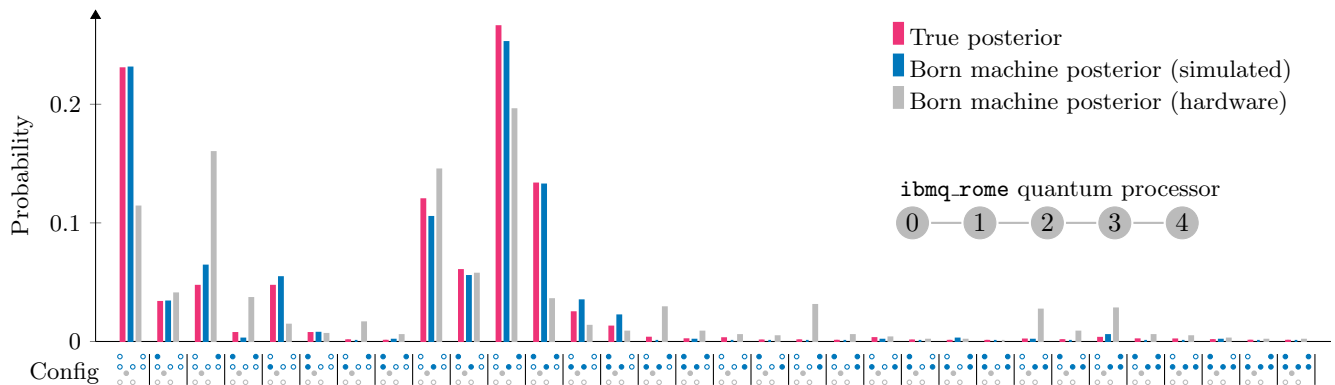
Figure 8: Histograms of true versus the learned posteriors with a simulated and hardware trained Born machine for the 'lung cancer' network in Fig. 1c. Here we use the `ibmq_rome` quantum processor (whose connectivity is shown in the inset). We condition on $X = \mathsf{fa}, D = \mathsf{fa}, I = \mathsf{tr}$. The x-axis shows the configuration of observed (gray: X, D, I) and unobserved variables (blue: A, S, T, L, B) corresponding to each probability (filled circles = tr, empty circles = fa). For hardware results, the histogram is generated using 1024 samples.

replacing the classical model in Fig. 2 with a quantum one). This generalization can be thought of as a quantum circuit version of a variational autoencoder [58, 95] and will be the focus of future work.

We also presented a second VI method based on the kernelized Stein discrepancy (KSD). A limitation of this approach is that it requires explicit priors and likelihoods. On the other hand, it provides plenty of flexibility in the choice of kernel. We used a generic Hamming kernel to compute similarities between bit-strings. For VI on graphical models, a graph-based kernel [115] that takes into account the structure could yield improved results. Moreover, one could attempt a demonstration of quantum advantage in VI using classically intractable kernels arising from quantum circuits [38–40].

Another interesting extension could be approximating the posterior with a Born machine with traced out additional qubits, i.e. locally purified states. This may allow trading a larger number of qubits for reduced circuit depth at constant expressivity [83].

We presented a few examples with direct application to financial, medical, and other domains. Another area of potential application of our methods is natural language processing. Quantum circuits have been proposed as a way to encode and compose words to form meanings [116–118]. Question answering with these models could be phrased as variational inference on quantum computers where our methods fit naturally.

Appendix A: Gradients for the adversarial method

In the adversarial VI method We optimize the objective functions Eqs. (4) and (7) via gradient ascent and descent, respectively. Assuming the optimal discriminator $d_{\boldsymbol{\phi}} = d^*$ has been found, the objective Eq. (7) becomes

$$\mathbb{E}_{\boldsymbol{x} \sim p_{\mathcal{D}}(\boldsymbol{x})} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})} \big[\mathrm{logit}(d^*(\boldsymbol{x}, \boldsymbol{z})) - \log p(\boldsymbol{x}|\boldsymbol{z}) \big]. \tag{A1}$$

While d^* implicitly depends on $\boldsymbol{\theta}$, the gradient of its expectation vanishes. To see this, recall that $\mathrm{logit}(d^*(\boldsymbol{z}, \boldsymbol{x})) = \log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) - \log p(\boldsymbol{z})$. Fixing $\boldsymbol{x}$ and taking the partial derivative for the j-th parameter:

$$\partial_{\theta_j} \sum_{\boldsymbol{z}} q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \sum_{\boldsymbol{z}} \partial_{\theta_j} \Big(q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \Big) \log(q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})) + \sum_{\boldsymbol{z}} q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \partial_{\theta_j} \Big(\log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \Big). \tag{A2}$$

The second term is zero since:

$$\sum_{\boldsymbol{z}} q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \partial_{\theta_j} \Big(\log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \Big) = \sum_{\boldsymbol{z}} \partial_{\theta_j} q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \partial_{\theta_j} \sum_{\boldsymbol{z}} q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \partial_{\theta_j} 1 = 0. \tag{A3}$$

The approximate posterior in Eq. (A2) is given by the expectations of observables $O_{\boldsymbol{z}} = |z\rangle\langle z|$, i.e. $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \langle O_{\boldsymbol{z}}\rangle_{\boldsymbol{\theta},\boldsymbol{x}}$. For a circuit parameterized by gates $U(\theta_j) = \exp\left(-i\frac{\theta_j}{2}H_j\right)$, where the Hermitian generator H_j has eigenvalues ± 1 (e.g. single-qubit rotations), the derivative of the first term in Eq. (A2) can be evaluated with the parameter shift rule [119–121]:

$$\partial_{\theta_j} q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \frac{q_{\boldsymbol{\theta}_j^+}(\boldsymbol{z}|\boldsymbol{x}) - q_{\boldsymbol{\theta}_j^-}(\boldsymbol{z}|\boldsymbol{x})}{2}, \tag{A4}$$

with $\boldsymbol{\theta}_j^{\pm} = \boldsymbol{\theta} \pm \frac{\pi}{2}\boldsymbol{e}_j$ and $\boldsymbol{e}_j$ a unit vector in the j-th direction. In summary, the partial derivative of the objective with respect to θ_j given the optimal discriminator is:

$$\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}(\boldsymbol{x})}\left[\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}_j^+}(\boldsymbol{z}|\boldsymbol{x})}\left[\mathrm{logit}(d^*(\boldsymbol{x},\boldsymbol{z})) - \log p(\boldsymbol{x}|\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\theta}_j^-}(\boldsymbol{z}|\boldsymbol{x})}\left[\mathrm{logit}(d^*(\boldsymbol{x},\boldsymbol{z})) - \log p(\boldsymbol{x}|\boldsymbol{z})\right]\right]. \tag{A5}$$

Appendix B: The Stein operator

Here we prove that $(O^p f)(\boldsymbol{z}) = s_p(\boldsymbol{x},\boldsymbol{z})^T f(\boldsymbol{z}) - \mathrm{tr}(\Delta f(\boldsymbol{z}))$ is a Stein operator for any function $f \colon \{0,1\}^n \to \mathbb{R}^n$ and probability mass function $p(\boldsymbol{z}|\boldsymbol{x})$ on $\{0,1\}^n$. To do so, we need to show that its expectation under the true posterior vanishes. Taking the expectation w.r.t. the true posterior we have:

$$\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})}[(O^p f)(\boldsymbol{z})] = \sum_{\boldsymbol{z}\in\{0,1\}^n} p(\boldsymbol{z}|\boldsymbol{x})\frac{\Delta p(\boldsymbol{x},\boldsymbol{z})^T}{p(\boldsymbol{x},\boldsymbol{z})}f(\boldsymbol{z}) - \sum_{\boldsymbol{z}\in\{0,1\}^n} p(\boldsymbol{z}|\boldsymbol{x})\,\mathrm{tr}(\Delta f(\boldsymbol{z})) \tag{B1}$$

The first summation is:

$$\sum_{\boldsymbol{z}\in\{0,1\}^n} \frac{1}{p(\boldsymbol{x})}\Delta p(\boldsymbol{x},\boldsymbol{z})^T f(\boldsymbol{z}) = \sum_{\boldsymbol{z}\in\{0,1\}^n}\sum_{i=1}^n (p(\boldsymbol{z}|\boldsymbol{x}) - p(\neg_i\boldsymbol{z}|\boldsymbol{x}))f_i(\boldsymbol{z}), \tag{B2}$$

while the second summation is:

$$\sum_{\boldsymbol{z}\in\{0,1\}^n} p(\boldsymbol{z}|\boldsymbol{x})\,\mathrm{tr}(\Delta f(\boldsymbol{z})) = \sum_{\boldsymbol{z}\in\{0,1\}^n}\sum_{i=1}^n p(\boldsymbol{z}|\boldsymbol{x})(f_i(\boldsymbol{z}) - f_i(\neg_i\boldsymbol{z})). \tag{B3}$$

Subtracting (B3) from (B2) we have:

$$\sum_{i=1}^n\left[\sum_{\boldsymbol{z}\in\{0,1\}^n} p(\boldsymbol{z}|\boldsymbol{x})f_i(\boldsymbol{z}) - \sum_{\boldsymbol{z}\in\{0,1\}^n} p(\neg_i\boldsymbol{z}|\boldsymbol{x})f_i(\boldsymbol{z}) - \sum_{\boldsymbol{z}\in\{0,1\}^n} p(\boldsymbol{z}|\boldsymbol{x})f_i(\boldsymbol{z}) + \sum_{\boldsymbol{z}\in\{0,1\}^n} p(\boldsymbol{z}|\boldsymbol{x})f_i(\neg_i\boldsymbol{z})\right] = 0. \tag{B4}$$

The first and third terms cancel. The second and fourth terms are identical and cancel also. This can be seen by substituting $\boldsymbol{y}_i = \neg_i\boldsymbol{z}$ in one of the summations and noting that $\boldsymbol{z} = \neg_i\boldsymbol{y}_i$. Thus the operator O^p satisfies Stein's identity $\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})}[(O^p f)(\boldsymbol{z})] = 0$ for any function f as defined above, and is a valid Stein operator.

Appendix C: Gradients for the kernelized method

We wish to optimize the kernelized Stein objective Eq. (14) via gradient descent. This gradient was derived in [42] and is given by:

$$\partial_{\theta_j}\mathcal{L}_{\mathrm{KSD}}(\boldsymbol{\theta}) = \partial_{\theta_j}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}(\boldsymbol{x})}\sqrt{\mathbb{E}_{\boldsymbol{z},\boldsymbol{z}'\sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')]}$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}(\boldsymbol{x})}\partial_{\theta_j}\sqrt{\mathbb{E}_{\boldsymbol{z},\boldsymbol{z}'\sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')]}$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}(\boldsymbol{x})}\left(\left\{\mathbb{E}_{\boldsymbol{z},\boldsymbol{z}'\sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')]\right\}^{-1/2}\partial_{\theta_j}\left[\mathbb{E}_{\boldsymbol{z},\boldsymbol{z}'\sim q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')]\right]\right)$$

$$= \frac{1}{4}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}(\boldsymbol{x})}\left[\frac{1}{\mathrm{KSD}\left[q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}|\boldsymbol{x})\right]}\left\{\mathbb{E}_{\substack{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}\\\boldsymbol{z}'\sim q_{\boldsymbol{\theta}_j^-}}}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')] - \mathbb{E}_{\substack{\boldsymbol{z}\sim q_{\boldsymbol{\theta}}\\\boldsymbol{z}'\sim q_{\boldsymbol{\theta}_j^+}}}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')] + \mathbb{E}_{\substack{\boldsymbol{z}\sim q_{\boldsymbol{\theta}_j^+}\\\boldsymbol{z}'\sim q_{\boldsymbol{\theta}}}}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')] - \mathbb{E}_{\substack{\boldsymbol{z}\sim q_{\boldsymbol{\theta}_j^-}\\\boldsymbol{z}'\sim q_{\boldsymbol{\theta}}}}[\kappa_p(\boldsymbol{z},\boldsymbol{z}')]\right\}\right] \tag{C1}$$

The final line (suppressing the dependence on the variables, z, x: $q_\theta := q_\theta(z|x)$) in Eq. (C1) is given directly by [42] and using the parameter shift rule, Eq. (A4) for the gradient of the output probabilities of the Born machine circuit, under the same assumptions as in Appendix A. The relatively simple form of the gradient is due to the fact that the Stein kernel, κ_p, does not depend on the variational distribution, $q_\theta(z|x)$. This gradient can be estimated in a similar way to the KSD itself, by sampling from the original Born machine, $q_\theta(z|x)$ along with its parameter shifted versions, $q_{\theta_j^\pm}(z|x)$, for each parameter, j.

Appendix D: Learning curves for the adversarial method

In this section we provide examples of successful and unsuccessful VI using the KL objective and adversarial method. We inspect the random instances used in the 'sprinkler' network experiment in section V A. We cherry-pick two out of the 30 instances where a 1-layer Born machine is employed. Figure (a) shows a successful experiment. Here the MLP classifier closely tracks the ideal classifier providing a good signal for the Born machine. The Born machine is able to minimize its loss, which in turns leads to ≈ 0 total variation distance. Figure (b) shows an unsuccessful experiment. Here, the MLP tracks the ideal classifier, but the Born machine is not able to find a 'direction' that minimizes the loss (see epochs 400 to 1000). We verified that a more powerful 2-layers Born machine performs much better for this instance (not shown).

In these Figures gray lines correspond to quantities that can be estimated from samples, while blue and magenta lines represent exact quantities that are not practical to compute. It is challenging to assess the training looking at the gray lines. This discussion is to emphasize the need for new techniques to assess the models. These would provide an early stopping criterion for unsuccessful experiments as well as a validation criterion for successful ones. This is a common problem among adversarial approaches.

(a) Successful example   **(b)** Unsuccessful example
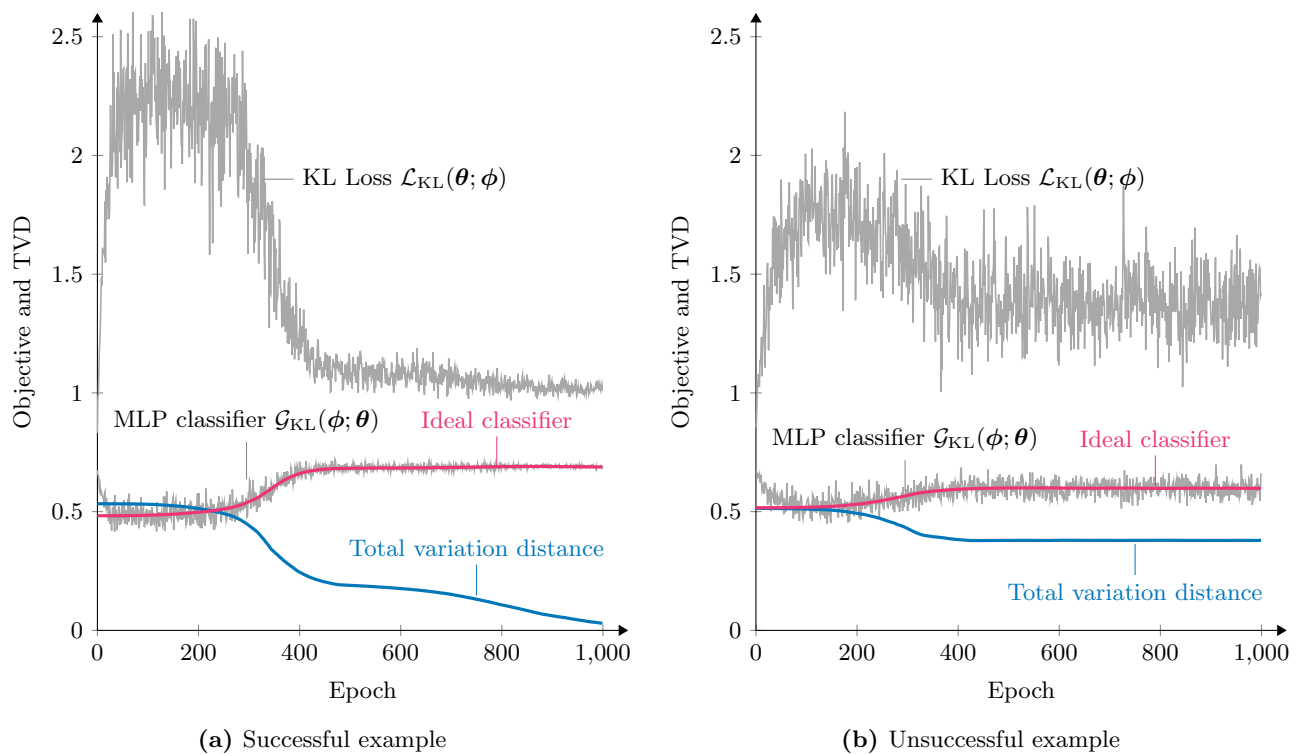
Figure 9: Examples of successful (a) and unsuccessful (b) VI on the 'sprinkler' Bayesian network in Sec. V A using the KL objective with the adversarial method.

Appendix E: Probability table for the 'lung cancer' network

The original 'lung cancer' network [49] contains a variable 'either' (meaning either lung cancer or tuberculosis are present). This node is deterministic and yields zero probabilities in some posterior distributions. In order to avoid numerical issues arising from zero probabilities, we replace 'either' by 'illness' which may be true even if both parent nodes 'lung cancer' and 'tuberculosis' are false. Table I shows the modified probability table for the network in Fig. 1c.

Variable	Probabilities		Variable	Probabilities
Asia (A)	$p(A = \mathsf{tr}) = 0.01$		Smoking (S)	$p(S = \mathsf{tr}) = 0.5$
	$p(A = \mathsf{fa}) = 0.99$			$p(S = \mathsf{fa}) = 0.5$
Tuberculosis (T)	$p(T = \mathsf{tr} \,|\, A = \mathsf{tr}) = 0.05$		Illness (I)	$p(I = \mathsf{tr} \,|\, L = \mathsf{tr}, T = \mathsf{tr}) = 1$
	$p(T = \mathsf{tr} \,|\, A = \mathsf{fa}) = 0.01$			$p(I = \mathsf{tr} \,|\, L = \mathsf{tr}, T = \mathsf{fa}) = 1$
	$p(T = \mathsf{fa} \,|\, A = \mathsf{tr}) = 0.95$			$p(I = \mathsf{tr} \,|\, L = \mathsf{fa}, T = \mathsf{tr}) = 1$
	$p(T = \mathsf{fa} \,|\, A = \mathsf{fa}) = 0.99$			$p(I = \mathsf{tr} \,|\, L = \mathsf{fa}, T = \mathsf{fa}) = 0.05$
Lung cancer (L)	$p(L = \mathsf{tr} \,|\, S = \mathsf{tr}) = 0.1$			$p(I = \mathsf{fa} \,|\, L = \mathsf{tr}, T = \mathsf{tr}) = 0$
	$p(L = \mathsf{tr} \,|\, S = \mathsf{fa}) = 0.01$			$p(I = \mathsf{fa} \,|\, L = \mathsf{tr}, T = \mathsf{fa}) = 0$
	$p(L = \mathsf{fa} \,|\, S = \mathsf{tr}) = 0.9$			$p(I = \mathsf{fa} \,|\, L = \mathsf{fa}, T = \mathsf{tr}) = 0$
	$p(L = \mathsf{fa} \,|\, S = \mathsf{fa}) = 0.99$			$p(I = \mathsf{fa} \,|\, L = \mathsf{fa}, T = \mathsf{fa}) = 0.95$
Bronchitis (B)	$p(B = \mathsf{tr} \,|\, S = \mathsf{tr}) = 0.6$		Dyspnoea (D)	$p(D = \mathsf{tr} \,|\, B = \mathsf{tr}, I = \mathsf{tr}) = 0.9$
	$p(B = \mathsf{tr} \,|\, S = \mathsf{fa}) = 0.3$			$p(D = \mathsf{tr} \,|\, B = \mathsf{tr}, I = \mathsf{fa}) = 0.8$
	$p(B = \mathsf{fa} \,|\, S = \mathsf{tr}) = 0.4$			$p(D = \mathsf{tr} \,|\, B = \mathsf{fa}, I = \mathsf{tr}) = 0.7$
	$p(B = \mathsf{fa} \,|\, S = \mathsf{fa}) = 0.7$			$p(D = \mathsf{tr} \,|\, B = \mathsf{fa}, I = \mathsf{fa}) = 0.1$
X-ray (X)	$p(X = \mathsf{tr} \,|\, I = \mathsf{tr}) = 0.98$			$p(D = \mathsf{fa} \,|\, B = \mathsf{tr}, I = \mathsf{tr}) = 0.1$
	$p(X = \mathsf{tr} \,|\, I = \mathsf{fa}) = 0.05$			$p(D = \mathsf{fa} \,|\, B = \mathsf{tr}, I = \mathsf{fa}) = 0.2$
	$p(X = \mathsf{fa} \,|\, I = \mathsf{tr}) = 0.02$			$p(D = \mathsf{fa} \,|\, B = \mathsf{fa}, I = \mathsf{tr}) = 0.3$
	$p(X = \mathsf{fa} \,|\, I = \mathsf{fa}) = 0.95$			$p(D = \mathsf{fa} \,|\, B = \mathsf{fa}, I = \mathsf{fa}) = 0.9$

Table I: Probability table corresponding to the network in Fig. 1c.

[1] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
[2] Q. Morris, Recognition networks for approximate inference in BN20 networks, in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI2001)*, edited by J. Breese and D. Koller (Morgan Kaufmann Publishers Inc., Seattle, Washington, USA, 2001) pp. 370–377, arXiv:1301.2295.
[3] J. G. Richens, C. M. Lee, and S. Johri, Improving the accuracy of medical diagnosis with causal machine learning, Nature Communications **11**, 3923 (2020).
[4] M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, *Handbook of Graphical Models*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods (CRC Press, Boca Raton, 2018).
[5] W. Wiegerinck, B. Kappen, and W. Burgers, Bayesian networks for expert systems: Theory and practical applications, in *Interactive Collaborative Information Systems*, edited by R. Babuška and F. C. A. Groen (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010) pp. 547–578.
[6] A. Denev, *Probabilistic Graphical Models: A New Way of Thinking in Financial Modelling* (Risk Books, 2015).
[7] B. Cai, L. Huang, and M. Xie, Bayesian Networks in Fault Diagnosis, IEEE Transactions on Industrial Informatics **13**, 2227 (2017).
[8] R. M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1 (Department of Computer Science, University of Toronto, Toronto, 1993).
[9] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds., *Handbook of Markov Chain Monte Carlo Methods* (Chapman & Hall/CRC, New York, NY, USA, 2011).
[10] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, Variational Inference: A Review for Statisticians, Journal of the American Statistical Association **112**, 859 (2017).
[11] S. H. Adachi and M. P. Henderson, Application of quantum annealing to training of deep neural networks (2015), arXiv:1510.06356 [quant-ph].
[12] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning, Phys. Rev. A **94**, 022308 (2016).

[13] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Quantum-assisted learning of hardware-embedded probabilistic graphical models, Phys. Rev. X **7**, 041052 (2017).

[14] D. Korenkevych, Y. Xue, Z. Bian, F. Chudak, W. G. Macready, J. Rolfe, and E. Andriyash, Benchmarking quantum hardware for training of fully visible Boltzmann machines (2016), arXiv:1611.04528 [quant-ph].

[15] M. Benedetti, J. Realpe-Gómez, and A. Perdomo-Ortiz, Quantum-assisted Helmholtz machines: A quantum–classical deep learning framework for industrial datasets in near-term devices, Quantum Science and Technology **3**, 034007 (2018).

[16] M. Wilson, T. Vandal, T. Hogg, and E. Rieffel, Quantum-assisted associative adversarial network: Applying quantum annealing in deep learning (2019), arXiv:1904.10573 [cs.LG].

[17] G. H. Low, T. J. Yoder, and I. L. Chuang, Quantum inference on Bayesian networks, Phys. Rev. A **89**, 062315 (2014).

[18] A. Montanaro, Quantum speedup of Monte Carlo methods, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **471**, 20150301 (2015).

[19] A. N. Chowdhury and R. D. Somma, Quantum algorithms for Gibbs sampling and hitting-time estimation, Quant. Inf. Comp. **17**, 0041 (2017), arXiv:1603.02940 [quant-ph].

[20] P. Wittek and C. Gogolin, Quantum Enhanced Inference in Markov Logic Networks, Scientific Reports **7** (2017).

[21] J. Wu and T. H. Hsieh, Variational thermal quantum simulation via thermofield double states, Phys. Rev. Lett. **123**, 220502 (2019).

[22] G. Verdon, J. Marks, S. Nanda, S. Leichenauer, and J. Hidary, Quantum Hamiltonian-Based Models and the Variational Quantum Thermalizer Algorithm (2019), arXiv:1910.02071 [quant-ph].

[23] A. N. Chowdhury, G. H. Low, and N. Wiebe, A Variational Quantum Algorithm for Preparing Quantum Gibbs States (2020), arXiv:2002.00055 [quant-ph].

[24] Y. Wang, G. Li, and X. Wang, Variational quantum Gibbs state preparation with a truncated Taylor series (2020), arXiv:2005.08797 [quant-ph].

[25] Y. Shingu, Y. Seki, S. Watabe, S. Endo, Y. Matsuzaki, S. Kawabata, T. Nikuni, and H. Hakoshima, Boltzmann machine learning with a variational quantum algorithm (2020), arXiv:2007.00876 [quant-ph].

[26] I. Sato, K. Kurihara, S. Tanaka, H. Nakagawa, and S. Miyashita, Quantum annealing for variational Bayes inference, in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09 (AUAI Press, Arlington, Virginia, USA, 2009) pp. 479–486, arXiv:1408.2037.

[27] B. O'Gorman, R. Babbush, A. Perdomo-Ortiz, A. Aspuru-Guzik, and V. Smelyanskiy, Bayesian network structure learning using quantum annealing, The European Physical Journal Special Topics **224**, 163–188 (2015).

[28] H. Miyahara and Y. Sughiyama, A Quantum Extension of Variational Bayes Inference, Physical Review A **98**, 022330 (2018), arXiv:1712.04709.

[29] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, Advances in variational inference (2018), arXiv:1711.05597 [cs.LG].

[30] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks (2020), arXiv:2011.00027 [quant-ph].

[31] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning (2021), arXiv:2101.02464 [quant-ph].

[32] K. Poland, K. Beer, and T. J. Osborne, No free lunch for quantum machine learning (2020), arXiv:2003.14103 [quant-ph].

[33] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles, Reformulation of the no-free-lunch theorem for entangled data sets (2020), arXiv:2007.04900 [quant-ph].

[34] R. A. Servedio and S. J. Gortler, Equivalences and Separations Between Quantum and Classical Learnability, SIAM Journal on Computing **33**, 1067 (2004).

[35] S. Arunachalam and R. de Wolf, Guest Column: A Survey of Quantum Learning Theory, ACM SIGACT News **48**, 41 (2017).

[36] S. Arunachalam and R. d. Wolf, Optimal Quantum Sample Complexity of Learning Algorithms, Journal of Machine Learning Research **19**, 1 (2018).

[37] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning (2020), arXiv:2010.02174 [quant-ph].

[38] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature **567**, 209 (2019).

[39] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[40] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning (2020), arXiv:2011.01938 [quant-ph].

[41] X. Gao, Z.-Y. Zhang, and L.-M. Duan, A quantum machine learning algorithm based on generative models, Science Advances **4**, eaat9004 (2018).

[42] B. Coyle, D. Mills, V. Danos, and E. Kashefi, The Born supremacy: quantum advantage and training of an Ising Born machine, npj Quantum Information **6** (2020).

[43] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, On the Quantum versus Classical Learnability of Discrete Distributions, Quantum **5**, 417 (2021).

[44] X. Gao, E. R. Anschuetz, S.-T. Wang, J. I. Cirac, and M. D. Lukin, Enhancing Generative Models via Quantum Correlations (2021), arXiv:2101.08354 [quant-ph].

[45] D. Ristè, M. P. da Silva, C. A. Ryan, A. W. Cross, A. D. Córcoles, J. A. Smolin, J. M. Gambetta, J. M. Chow, and B. R. Johnson, Demonstration of quantum advantage in machine learning, npj Quantum Information **3**, 1 (2017).

[46] B. Coyle, M. Henderson, J. C. J. Le, N. Kumar, M. Paini, and E. Kashefi, Quantum versus classical generative modelling in finance, Quantum Science and Technology 10.1088/2058-9565/abd3db (2020).

[47] J. Alcazar, V. Leyton-Ortega, and A. Perdomo-Ortiz, Classical versus quantum models in machine learning: insights from a finance application, Machine Learning: Science and Technology **1**, 035003 (2020).

[48] S. Johri, S. Debnath, A. Mocherla, A. Singh, A. Prakash, J. Kim, and I. Kerenidis, Nearest centroid classification on a trapped ion quantum computer (2020), arXiv:2012.04145 [quant-ph].

[49] S. L. Lauritzen and D. J. Spiegelhalter, Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems, Journal of the Royal Statistical Society. Series B (Methodological) **50**, 157 (1988).

[50] M. Kritzman, S. Page, and D. Turkington, Regime Shifts: Implications for Dynamic Strategies (corrected), Financial Analysts Journal **68**, 22 (2012).

[51] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science **366**, 447 (2019).

[52] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. (Cambridge University Press, New York, NY, USA, 2009).

[53] R. McEliece, D. MacKay, and Jung-Fu Cheng, Turbo decoding as an instance of Pearl's "belief propagation" algorithm, IEEE Journal on Selected Areas in Communications **16**, 140 (1998).

[54] Dan Roth, On the hardness of approximate reasoning, Artificial Intelligence **82**, 273 (1996).

[55] G. F. Cooper, The computational complexity of probabilistic inference using bayesian belief networks, Artificial Intelligence **42**, 393 (1990).

[56] P. Dagum and M. Luby, Approximating probabilistic inference in Bayesian belief networks is NP-hard, Artificial Intelligence **60**, 141 (1993).

[57] R. Ranganath, S. Gerrish, and D. M. Blei, Black box variational inference (2013), arXiv:1401.0118 [stat.ML].

[58] D. P. Kingma and M. Welling, Auto-encoding variational Bayes (2014), arXiv:1312.6114 [stat.ML].

[59] D. Wingate and T. Weber, Automated variational inference in probabilistic programming (2013), arXiv:1301.1299 [stat.ML].

[60] S. Gershman, M. Hoffman, and D. Blei, Nonparametric variational inference (2012), arXiv:1206.4665 [cs.LG].

[61] D. J. Rezende and S. Mohamed, Variational inference with normalizing flows (2016), arXiv:1505.05770 [stat.ML].

[62] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine learning **8**, 229 (1992).

[63] C. J. Maddison, A. Mnih, and Y. W. Teh, The concrete distribution: A continuous relaxation of discrete random variables (2017), arXiv:1611.00712 [cs.LG].

[64] E. Jang, S. Gu, and B. Poole, Categorical reparameterization with gumbel-softmax (2017), arXiv:1611.01144 [stat.ML].

[65] S. Cheng, J. Chen, and L. Wang, Information perspective to probabilistic modeling: Boltzmann machines versus Born machines, Entropy **20**, 583 (2018).

[66] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, A generative modeling approach for benchmarking and training shallow quantum circuits, npj Quantum Information **5** (2019).

[67] K. E. Hamilton, E. F. Dumitrescu, and R. C. Pooser, Generative model benchmarks for superconducting qubits, Phys. Rev. A **99**, 062323 (2019).

[68] V. Leyton-Ortega, A. Perdomo-Ortiz, and O. Perdomo, Robust implementation of generative modeling with parametrized quantum circuits (2019), arXiv:1901.08047 [quant-ph].

[69] D. Zhu, N. M. Linke, M. Benedetti, K. A. Landsman, N. H. Nguyen, C. H. Alderete, A. Perdomo-Ortiz, N. Korda, A. Garfoot, C. Brecque, L. Egan, O. Perdomo, and C. Monroe, Training of quantum circuits on a hybrid quantum computer, Science Advances **5** (2019).

[70] Z.-Y. Han, J. Wang, H. Fan, L. Wang, and P. Zhang, Unsupervised generative modeling using matrix product states, Phys. Rev. X **8**, 031012 (2018).

[71] M. S. Rudolph, N. B. Toussaint, A. Katabarwa, S. Johri, B. Peropadre, and A. Perdomo-Ortiz, Generation of high-resolution handwritten digits with an ion-trap quantum computer (2020), arXiv:2012.03924 [quant-ph].

[72] I. Čepaitė, B. Coyle, and E. Kashefi, A Continuous Variable Born Machine (2020), arXiv:2011.00904 [quant-ph].

[73] C. Zoufal, A. Lucchi, and S. Woerner, Quantum generative adversarial networks for learning and loading random distributions, npj Quantum Information **5** (2019).

[74] D. Herr, B. Obert, and M. Rosenkranz, Anomaly detection with variational quantum generative adversarial networks (2020), arXiv:2010.10492 [quant-ph].

[75] P.-L. Dallaire-Demers and N. Killoran, Quantum generative adversarial networks, Phys. Rev. A **98**, 012324 (2018).

[76] M. Benedetti, E. Grant, L. Wossnig, and S. Severini, Adversarial quantum circuit learning for pure state approximation, New Journal of Physics **21**, 043023 (2019).

[77] S. Chakrabarti, H. Yiming, T. Li, S. Feizi, and X. Wu, Quantum Wasserstein generative adversarial networks, in *Advances in Neural Information Processing Systems* (2019) pp. 6781–6792.

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

[79] S. Aaronson and A. Arkhipov, The Computational Complexity of Linear Optics, in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11 (Association for Computing Machinery, New York, NY, USA, 2011) p. 333–342.

[80] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Characterizing quantum supremacy in near-term devices, Nature Physics **14**, 595 (2018).

[81] A. Bouland, B. Fefferman, C. Nirkhe, and U. Vazirani, On the complexity and verification of quantum random circuit sampling, Nature Physics **15**, 159 (2019).

[82] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, Expressive power of parametrized quantum circuits, Physical Review Research **2** (2020).

[83] I. Glasser, R. Sweke, N. Pancotti, J. Eisert, and J. I. Cirac, Expressive power of tensor-network factorizations for probabilistic modeling, with applications from hidden Markov models to quantum machine learning (2019), arXiv:1907.03741 [cs.LG].

[84] S. Mohamed and B. Lakshminarayanan, Learning in implicit generative models (2017), arXiv:1610.03483 [stat.ML].

[85] S. Aaronson, Quantum computing, postselection, and probabilistic polynomial-time, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **461**, 3473 (2005).

[86] G. D. las Cuevas, W. Dür, M. V. den Nest, and M. A. Martin-Delgado, Quantum algorithms for classical lattice models, New Journal of Physics **13**, 093021 (2011).

[87] K. Fujii and T. Morimae, Commuting quantum circuits and complexity of Ising partition functions, New Journal of Physics **19**, 033003 (2017).

[88] J.-G. Liu and L. Wang, Differentiable learning of quantum circuit Born machines, Phys. Rev. A **98**, 062324 (2018).

[89] H. Situ, Z. He, Y. Wang, L. Li, and S. Zheng, Quantum generative adversarial network for generating discrete distribution, Information Sciences **538**, 193–208 (2020).

[90] J. Zeng, Y. Wu, J.-G. Liu, L. Wang, and J. Hu, Learning and inference on generative adversarial quantum circuits, Phys. Rev. A **99**, 052306 (2019).

[91] S. E. Borujeni, S. Nannapaneni, N. H. Nguyen, E. C. Behrman, and J. E. Steck, Quantum circuit representation of Bayesian networks (2020), arXiv:2004.14803 [quant-ph].

[92] R. R. Tucci, Quantum Bayesian Nets, International Journal of Modern Physics B **09**, 295 (1995), arXiv:quant-ph/9706039.

[93] R. Ranganath, J. Altosaar, D. Tran, and D. M. Blei, Operator variational inference (2018), arXiv:1610.09033 [stat.ML].

[94] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, On integral probability metrics, ϕ-divergences and binary classification (2009), arXiv:0901.2698 [cs.IT].

[95] D. J. Rezende, S. Mohamed, and D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models (2014), arXiv:1401.4082 [stat.ML].

[96] L. Mescheder, S. Nowozin, and A. Geiger, Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks (2018), arXiv:1701.04722 [cs.LG].

[97] F. Huszár, Variational inference using implicit distributions (2017), arXiv:1702.08235 [stat.ML].

[98] J. Yang, Q. Liu, V. Rao, and J. Neville, Goodness-of-fit testing for discrete distributions via Stein discrepancy, in *International Conference on Machine Learning* (2018) pp. 5561–5570.

[99] J. Gorham and L. Mackey, Measuring Sample Quality with Stein's Method (2018), arXiv:1506.03039 [cs, math, stat].

[100] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nature Communications **9**, 4812 (2018).

[101] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nature Communications **12** (2021).

[102] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms (2021), arXiv:2007.14384 [quant-ph].

[103] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus (2021), arXiv:2101.02138 [quant-ph].

[104] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks (2020), arXiv:2011.02966 [quant-ph].

[105] C. Zhao and X.-S. Gao, Analyzing the barren plateau phenomenon in training quantum neural network with the zx-calculus (2021), arXiv:2102.01828 [quant-ph].

[106] R. LaRose and B. Coyle, Robust data encodings for quantum classifiers, Phys. Rev. A **102**, 032420 (2020).

[107] M. Schuld, R. Sweke, and J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, Phys. Rev. A **103**, 032430 (2021).

[108] Z. Ghahramani and M. I. Jordan, Factorial hidden Markov models, Machine learning **29**, 245 (1997).

[109] J. Z. Kolter and T. Jaakkola, Approximate inference in additive factorial hmms with application to energy disaggregation, in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 22, edited by N. D. Lawrence and M. Girolami (PMLR, La Palma, Canary Islands, 2012) pp. 1472–1482.

[110] R. S. Smith, M. J. Curtis, and W. J. Zeng, A practical quantum instruction set architecture (2016), arXiv:1608.03355 [quant-ph].

[111] S. Sivarajah, S. Dilkes, A. Cowtan, W. Simmons, A. Edgington, and R. Duncan, t|ket⟩: a retargetable compiler for NISQ devices, Quantum Science and Technology **6**, 014003 (2020).

[112] D. Barber, *Bayesian Reasoning and Machine Learning* (Cambridge University Press, Cambridge, UK, 2012).

[113] M. Arjovsky and L. Bottou, Towards principled methods for training generative adversarial networks (2017), arXiv:1701.04862 [stat.ML].

[114] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, Are GANs Created Equal? A Large-Scale Study, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18 (Curran Associates Inc., Red Hook, NY, USA, 2018) pp. 698–707, arXiv:1711.10337 [stat.ML].

[115] R. I. Kondor and J. D. Lafferty, Diffusion kernels on graphs and other discrete input spaces, in *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002) p. 315–322.

[116] B. Coecke, G. de Felice, K. Meichanetzidis, and A. Toumi, Foundations for near-term quantum natural language processing (2020), arXiv:2012.03755 [quant-ph].

[117] K. Meichanetzidis, A. Toumi, G. de Felice, and B. Coecke, Grammar-aware question-answering on quantum computers (2020), arXiv:2012.03756 [quant-ph].

[118] R. Lorenz, A. Pearson, K. Meichanetzidis, D. Kartsaklis, and B. Coecke, Qnlp in practice: Running compositional models of meaning on a quantum computer (2021), arXiv:2102.12846 [cs.CL].

[119] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, Phys. Rev. A **98**, 032309 (2018).

[120] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A **99**, 032331 (2019).

[121] L. Banchi and G. E. Crooks, Measuring Analytic Gradients of General Quantum Evolution with the Stochastic Parameter Shift Rule, Quantum **5**, 386 (2021).