

Learning energy-based representations of quantum many-body states

Abhijith Jayakumar^{1,2,*}, Marc Vuffray¹, and Andrey Y. Lokhov¹

¹Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

²Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA



(Received 5 July 2023; accepted 3 July 2024; published 21 August 2024)

Efficient representation of quantum many-body states on classical computers is a problem of practical importance. An ideal representation of a quantum state combines a succinct characterization informed by the structure and symmetries of the system along with the ability to predict the physical observables of interest. Several machine-learning approaches have been recently used to construct such classical representations, which enable predictions of observables and account for physical symmetries. However, the structure of a quantum state typically gets lost unless a specialized *Ansatz* is employed based on prior knowledge of the system. Moreover, most such approaches give no information about what states are easier to learn in comparison with others. Here, we propose a generative energy-based representation of quantum many-body states derived from Gibbs distributions used for modeling the thermal states of classical spin systems. Based on the prior information on a family of quantum states, the energy function can be specified by a small number of parameters using an explicit low-degree polynomial or a generic parametric family such as neural nets and can naturally include the known symmetries of the system. Our results show that such a representation can be efficiently learned from data using exact algorithms in a form that enables the prediction of expectation values of physical observables. Importantly, the structure of the learned energy function provides a natural explanation for the difficulty of learning an energy-based representation of a given class of quantum states when measured in a certain basis.

DOI: [10.1103/PhysRevResearch.6.033201](https://doi.org/10.1103/PhysRevResearch.6.033201)

I. INTRODUCTION

Learning a generative model from several copies of a quantum state is quickly becoming an important task due to the ever-increasing size of quantum systems that can be prepared using various quantum information processing devices [1,2]. The quantum state with n qubits is fully characterized by the density matrix $\rho \in \mathbb{C}^{2^n \times 2^n}$ with $2^{2n} - 1$ parameters. Due to the exponentially increasing number of parameters with the system size, directly learning a quantum density matrix becomes rapidly impractical [3,4] except in special cases, such as matrix product states (MPS) [5,6], where hand-tailored quantum tomography techniques exploiting the known structure of the state can improve the representation efficiency [5,7–9]. Alternatively, some compact representations for states can be hard to learn or can be inefficient for the purposes of estimating observables. For example, the thermal state of a local quantum Hamiltonian can be specified using polynomially few parameters that encode the structure of the state. However, this representation is not very useful in practice because, currently, no computationally efficient algorithms to learn the quantum Hamiltonian parameters are known, except in special cases [10–13]. Moreover, even if the quantum Hamiltonian is

known, estimating the observables is, in general, difficult due to issues like the sign problem [14]. It is worth mentioning that, if the aim of the tomographic technique is not to learn a generative model but to estimate a fixed subset of observables, then recent advances in *shadow tomography* can be used in many practically interesting settings [15,16].

In many cases, rather than trying to represent the quantum state directly, a more fruitful approach would be to learn a classical generative model for the measurement statistics generated by a quantum state. Since a state can be completely specified by the distribution of its measurement outcomes in an appropriately chosen basis, such a classical model can be used as an equivalent model for the quantum state. Most commonly, the mapping of a quantum state to a classical distribution is performed using multiqubit *positive operator valued measures* (POVMs) in the form of a tensor product of informationally complete single-qubit POVMs [17,18]. These POVMs are specified using operators \mathbf{M}_{σ_i} for each qubit i , where $\sigma_i \in \{1, \dots, q\}$ is a classical variable that takes one of the q states. For a quantum state on n qubits, such a mapping generates a classical n -body distribution on classical variable configurations $\underline{\sigma} = (\sigma_1, \dots, \sigma_n)$:

$$\mu(\underline{\sigma}) = \text{Tr}[\rho \mathbf{M}_{\sigma_1}^{(1)} \otimes \mathbf{M}_{\sigma_2}^{(2)} \otimes \dots \otimes \mathbf{M}_{\sigma_n}^{(n)}], \quad (1)$$

which acts as an unambiguous representation of the quantum state ρ . Noninformationally complete POVMs can also be used to produce a reduced set of measurements, in which case $\mu(\underline{\sigma})$ models only partial information on the quantum state. Given $\mu(\underline{\sigma})$, the expectation value of any observable with respect to ρ (or a reduced representation of ρ) can be computed if it is possible to sample from $\mu(\underline{\sigma})$. These ideas

*Contact author: abhijithj@lanl.gov

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

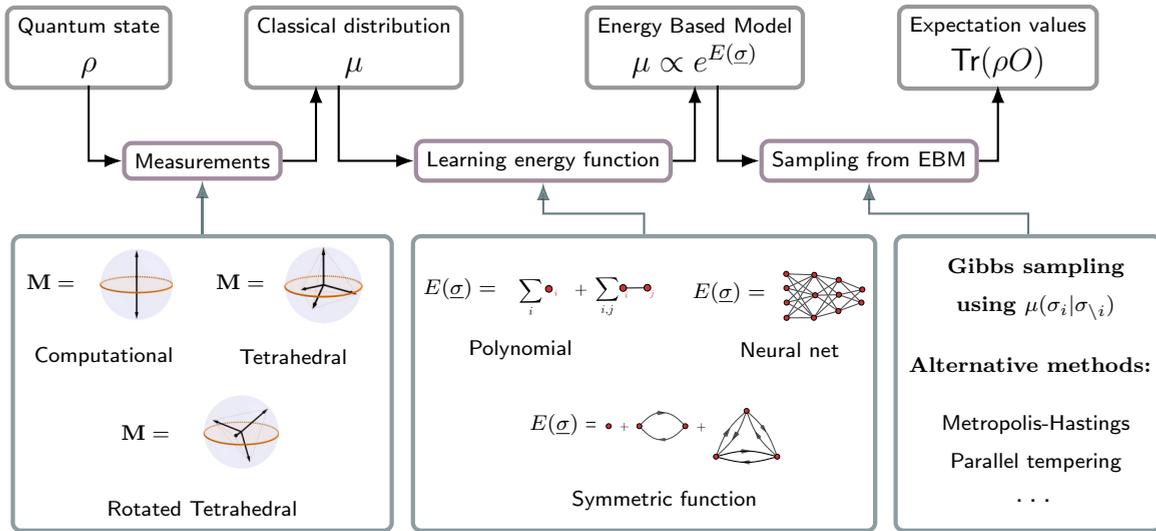


FIG. 1. Summary of our approach to learning of energy-based representations for quantum states. Our framework contains three modules in which different choices are possible. First, the choice of positive operator valued measure (POVM) defines a mapping from the quantum state to the corresponding classical representation. Second, the choice of the parametric family is related to the complexity of the energy function as well as to the choice of the respective learning algorithm from the family of interaction screening estimators. Finally, once the energy function is learned, a suitable sampling algorithm can be used to generate samples and estimate observation values using the conditional probabilities obtained during the learning process. As mentioned in the introduction, neural nets are advantageous to use when no prior information is available about the quantum state. Polynomials are advantageous to use if the quantum state exhibits a low-degree structure, as seen for the transverse Ising model. If we know *a priori* that the state is fully symmetric, then the symmetric function *Ansatz* is advantageous to use.

have been exploited in recent works, where machine-learning techniques originally developed to represent classical distributions have been used to learn representations for quantum many-body systems.

A popular approach in generative modeling of quantum states, introduced by Carleo and Troyer [19], is to model the amplitude and phases of a pure state using separate models. In an attempt to extend this representation to model practically abundant mixed states, recent works used different approaches such as introducing new degrees of freedom to purify the state [20], exploiting low-rank properties [21], or using matrix factorization [22]. These methods often lead to black-box models for quantum states which makes it difficult to impose prior information about the state, such as locality properties or symmetries. A natural approach to modeling mixed states was used by Carrasquilla *et al.* [17], who directly used measurement statistics from a quantum state to construct the respective probabilistic representation using POVMs. This approach allows for the direct use of generative modeling techniques for learning quantum states.

The major open question regarding learning of a classical distribution describing measurement data remains the lack of characterization of which states are hard or easy to model with standard generative modeling techniques. This makes choosing the right method for a given class of states an exercise in trial and error. Moreover, whereas many modern machine-learning approaches rely on the generalization properties of neural networks to find the sparsity structure in the probability density, such a structure may only be apparent at the level of the energy function, as it is, for instance, the case for the thermal states of local Hamiltonians. In this paper, we aim to rectify some of the common shortcomings encountered in a

direct application of machine-learning techniques to generative modeling of quantum states. We achieve this by modeling a distribution on statistics of measurement outcomes as an *energy-based model* (EBM):

$$\mu(\underline{\sigma}) = \frac{1}{Z} e^{E(\underline{\sigma})}, \tag{2}$$

and by focusing on learning the general real-valued energy function $E(\underline{\sigma})$ instead of the density itself. This approach is inspired by the power of Gibbs distributions used in modeling thermal states of classical many-body systems, where a simple quadratic energy function in the microscopic degrees of freedom induces distributions with highly nontrivial spin-glass structure [23]. In applications beyond modeling classical thermal systems, EBMs have seen a recent resurgence in the generative modeling literature due to their simplicity and state-of-the-art generalization abilities on real-world datasets [24–27]. In what follows, we leverage recent progress in rigorous learning of Gibbs distributions [28] to identify the energy function for various families of pure and mixed quantum states and to get insights into the complexity of representation, both from the perspective of learning and generating predictions, from the structure of the learned energy function and its effective temperature.

Our approach is schematically summarized in Fig. 1. Our aim is to learn an EBM for the distribution $\mu(\underline{\sigma})$, given m measurement outcomes (i.e., samples from μ) obtained by measuring independent copies of an n -qubit quantum state ρ using a particular POVM. The particular choice of the POVM determines the nature of the classical representation $\mu(\underline{\sigma})$. In this paper, we use computational, tetrahedral, and rotated tetrahedral POVMs, which are defined in the Methods

section. By studying the properties of $\mu(\sigma)$, we can make an appropriate choice for the parametric family to represent the energy function $E(\sigma)$. In the absence of any prior information, we use a generic neural net parametric family to discover a sparse nonpolynomial representation for the energy function [29]. When the structure is given by a low-order polynomial or some prior information on the symmetries of the state is available, we use more specialized representations such as polynomial or symmetric function families, respectively. Given a specific parametric family, we use a state-of-the-art computationally and sample-efficient method known as *interaction screening* (IS) [28–31] to learn the parameters of the energy function. This learning procedure bypasses the intractability of the maximum likelihood and provides access to the conditional probabilities $\mu(\sigma_i|\underline{\sigma}_{\setminus i})$ for each variable i in the EBM. These conditionals can then be used by a Markov chain Monte Carlo (MCMC) method to effectively generate samples from the learned EBM and to estimate any expectation value the POVM gives us access to [29,32]. In this paper, we use Gibbs sampling for estimating observables from the learned model, although other methods can be utilized depending on the application focus [33–35]. Our methodology naturally satisfies the two main desirable conditions for the generative modeling of quantum states: effective and practical algorithms for learning a classical representation and the ability to infer quantities of interest from that learned representation. Finally, the structural information from the inferred effective energy function, such as the intensity of parameters and sparsity of interactions, can be directly connected to the information-theoretic complexity of the resulting representation [36].

The main contributions of this paper are as follows. We find that, for many classes of quantum states, including thermal and ground states of local Hamiltonians, the choice of the correct function family to use to learn the EBM can be effectively made by studying systems on a small number of qubits. We can tractably learn these smaller systems in the infinite sample limit ($m \rightarrow \infty$), as explained in the Methods section. This helps us to understand the nature of the exact EBM representation for the state and use this knowledge to inform our choices for larger systems. For a given choice of the parametric function family modeling the energy function, we show how the effective temperature emerging from the learning procedure serves as a metric that quantifies the hardness of learning of different classes of quantum states. This is natural because the inverse temperature is the leading parameter that enters the information-theoretic bounds for learning of EBMs. Finally, we showcase the advantages of *Ansätze* that use a low number of parameters for representing quantum states. In all cases, we find that the EBM approach is well suited to learning classical generative models for quantum states and can be used to accurately estimate relevant observables. Our scaling experiments show that these energy-based methods are suitable for learning representations for large quantum systems, see Appendix E. As a particular example, we find that a low-parameter symmetric function *Ansatz* for learning the energy function helps us perform better in fidelity estimation tasks when compared with other neural-net-based methods [15].

Like any method that learns classical representation of quantum states, the methods outlined in this paper also

have limitations. While the aim of this paper is to give a comprehensive method for EBM representations of quantum states, note that such a representation might not be suitable for all states. For example, states which are sharply peaked in the measurement basis can produce EBMs with very low-temperature which are inherently hard to learn from information theoretic arguments [37]. In such a case, other methods for modeling the observed distribution might be advantageous. However, in this paper, we observe that many interesting classes of states do not possess such adversarial properties.

II. METHODS

In this section, we give a brief overview of the methods used in this paper.

A. POVMs and the probabilistic representation of quantum states

For constructing classical representations of quantum states, we use a probability distribution that is generated by the quantum state when we perform a certain set of measurements on it. We use a POVM to represent such a set of measurements [17,18,38]. POVMs are a set of positive operators such that they sum to the identity. A POVM with q elements is defined as

$$\mathbf{M}_\sigma \geq 0 \quad \forall \sigma \in [q], \quad \sum_{\sigma} \mathbf{M}_\sigma = I. \quad (3)$$

The POVM maps a quantum state ρ to a probability vector according to the Born rule:

$$\mu(\sigma) = \text{Tr}(\mathbf{M}_\sigma \rho). \quad (4)$$

Here, $\mu(\sigma)$ represents the probability of observing σ after measuring the state ρ in this POVM.

For many-body states whose Hilbert spaces have a tensor product structure, POVMs can be constructed by taking tensor products of single-body POVMs. For an n qubit system with a POVM of size q for each qubit, the POVM elements for the full system will have the form:

$$\mathbf{M}_{\underline{\sigma}} = \mathbf{M}_{\sigma_1}^{(1)} \otimes \mathbf{M}_{\sigma_2}^{(2)} \otimes \dots \otimes \mathbf{M}_{\sigma_n}^{(n)}. \quad (5)$$

Here, $\underline{\sigma}$ is now a multi-index that runs over all the q^n possibilities. This POVM naturally maps a quantum state over n qubits to a high-dimensional probability distribution:

$$\mu(\underline{\sigma}) = \text{Tr}(\mathbf{M}_{\underline{\sigma}} \rho). \quad (6)$$

Measuring the quantum state using the above POVM gives us samples drawn from this distribution. We can then use these samples to learn an EBM for the quantum state in question.

A POVM is known as *informationally complete* if the relation in Eq. (1) can be inverted to unambiguously reconstruct the quantum state ρ . This inversion can be performed using a set of dual operators defined by

$$\mathbf{D}_{\underline{\sigma}} = \sum_{\underline{\sigma}'} [C^{-1}]_{\underline{\sigma}, \underline{\sigma}'} \mathbf{M}_{\underline{\sigma}'}, \quad C_{\underline{\sigma}, \underline{\sigma}'} = \text{Tr}(\mathbf{M}_{\underline{\sigma}} \mathbf{M}_{\underline{\sigma}'}). \quad (7)$$

If POVM operators are linearly independent, then the matrix C is always invertible. On top of linear independence, if the

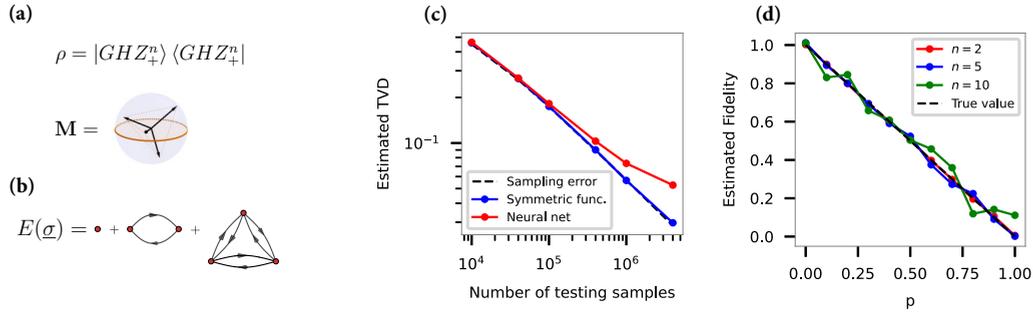


FIG. 2. Learning classical energy-based representations for permutation invariant systems. (a) We use the tetrahedral positive operator valued measure (POVM) rotated by a fixed 1 qubit Haar random unitary for learning Greenberger-Horne-Zeilinger state. (b) Given prior information about the symmetry of the state, we can use a fully permutation-invariant function in the interaction screening (IS) method. (c) The total variation distance (TVD) of the learned distribution from the true distribution obtained by measuring a 7 qubit GHZ state in the tetrahedral POVM. The x axis here is the number of samples drawn from a learned model to estimate the TVD. The dotted line is the sampling error, which is TVD estimated from the exact classical representation of the state given by the POVM. We see that the symmetric function *Ansatz* performs much better than the neural net *Ansatz* learned with NeurISE [29] without any symmetry assumptions. Both models are learned using 10^6 state measurements. The neural net has 1043 trainable parameters, while the symmetric function has 840 parameters. (d) Comparison of the fidelity between $\rho_{n,p}$ for $p \in [0, 1]$ with $|\text{GHZ}_+^n\rangle$ estimated from a symmetric energy-based model (EBM) for $\rho_{n,p}$. The symmetric EBM can produce the correct fidelity behavior here even when $p \rightarrow 1$. It has been reported that, for this benchmark, neural net quantum states (NNQS) methods which compute a classical upper bound for quantum fidelity cannot reproduce the correct behavior when p gets close to 1 [15, 17]. All the models here are learned using 6×10^4 samples.

POVM operators can span the entire operator space, then the POVM is informationally complete. In this case, the density matrix can be reconstructed from the distribution of measurement outcomes using the dual operators:

$$\rho = \sum_{\underline{\sigma}} \mu(\underline{\sigma}) \mathbf{D}_{\underline{\sigma}}. \quad (8)$$

If the POVM operators for an n -qubit system are formed by the tensor product of single-qubit POVMs, then both the C matrix and the dual operators will inherit this tensor product structure. This property allows for the estimation of local observables and other quantities of interest from a model that can generate samples from μ . If $\underline{\tau}^{(1)}, \dots, \underline{\tau}^{(N)}$ are such samples, then an unbiased estimate for $\text{Tr}(\rho O)$ can be constructed for any O :

$$\begin{aligned} \text{Tr}(\rho O) &= \mathbb{E}_{\underline{\tau} \sim \mu} \text{Tr}(\mathbf{D}_{\underline{\tau}} O) \\ &\approx \frac{1}{N} \sum_{k=1}^N \text{Tr}[\mathbf{D}_{\underline{\tau}^{(k)}} O] \\ &= \frac{1}{N} \sum_{k=1}^N \text{Tr}[O \mathbf{D}_{\tau_1^{(k)}}^{(1)} \otimes \mathbf{D}_{\tau_2^{(k)}}^{(2)} \otimes \dots \otimes \mathbf{D}_{\tau_n^{(k)}}^{(n)}]. \quad (9) \end{aligned}$$

If the operator O has an efficient representation as an matrix product operator (MPO) or as a sum of a few Pauli strings, then this estimate can be computed efficiently [17].

Most of the currently available qubit devices natively support measurements in the computational basis. This is a noninformationally complete POVM that corresponds to an EBM where each variable in the model can take two possible values (the number of possible outcomes of a measurement q gives the *alphabet size* of the EBM. In this case, the alphabet size is $q = 2$). The POVM operators corresponding to

computational basis measurements are given below:

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \quad (10)$$

We also use the tetrahedral POVM, which is one of the simplest informationally complete POVMs with $q = 4$. The single-qubit POVM operators for the tetrahedral POVM are

$$\begin{aligned} \mathbf{M}_1 &= \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}, \\ \mathbf{M}_2 &= \frac{1}{2} \begin{pmatrix} \frac{1}{3} & \frac{\sqrt{2}}{3} \\ \frac{\sqrt{2}}{3} & \frac{2}{3} \end{pmatrix}, \\ \mathbf{M}_3 &= \frac{1}{2} \begin{bmatrix} \frac{1}{3} & \frac{\sqrt{2}}{3} \exp(-i\frac{2\pi}{3}) \\ \frac{\sqrt{2}}{3} \exp(i\frac{2\pi}{3}) & \frac{2}{3} \end{bmatrix}, \\ \mathbf{M}_4 &= \frac{1}{2} \begin{bmatrix} \frac{1}{3} & \frac{\sqrt{2}}{3} \exp(-i\frac{4\pi}{3}) \\ \frac{\sqrt{2}}{3} \exp(i\frac{4\pi}{3}) & \frac{2}{3} \end{bmatrix}. \quad (11) \end{aligned}$$

We also use a rotated version of this POVM, where we rotate each of the POVM operators with a single Haar-random unitary, in the experiments with the Greenberger-Horne-Zeilinger state in Figs. 2(a) and 2(b). In our experiments with the GHZ state, we find that the rotated version of the tetrahedral POVM performs better than the unrotated version (see Appendix E). This points to the fact that difficulty of learning an EBM representation is also dependent on the choice of POVM.

In practice, such general POVMs are not usually natively supported in current quantum devices. For instance, IBM quantum computers only allow measurements in the computational basis, but measurement outcomes from other POVMs can be easily generated by using ancilla qubits and classical

postprocessing from such computational basis measurements [39,40].

B. Learning EBMs by IS

Like in any machine-learning problem, learning the energy function becomes prohibitively hard if we do not restrict the hypothesis space of possible energy functions. Hence, it is necessary to choose a parametric family of functions such that the learned model can be expressed as a member of this family, ideally using a small number of free parameters. In the case of representing quantum states, this parametric family can be chosen based on the prior information available about the state. In this paper, we will work with polynomials with a fixed degree, feed-forward neural nets, and symmetric functions as parametric families for representing the energy functions.

Learning an EBM then refers to the task of finding a function in this parametric family that accurately models the energy function in Eq. (2), given samples from the distribution μ . Standard parametric estimation techniques like maximum likelihood are inefficient for learning the energy function, except within a very restricted function class [41,42]. Instead, we use a computationally efficient and sample optimal approach known as IS [28–31] for learning the energy function. For each variable in the model, this method finds a representation of the part of the energy function connected to that variable. For example, for the EBM with an energy function $E(\underline{\sigma}) = \sum_{i,j} J_{ij}\sigma_i\sigma_j$, the IS method using a polynomial representation would learn the function $\sum_{j \neq i} J_{ij}\sigma_j$ for each i . We call this part of the energy function the *local energy* associated with variable σ_i .

For any EBM, given the local energy for any σ_i , the conditional of this variable given every other variable $\mu(\sigma_i|\underline{\sigma}_{\setminus i})$ can be easily expressed (see Appendix B for more details). These conditionals can then be used by a MCMC method to effectively generate samples from the learned EBM and to estimate any expectation value the POVM gives us access to [29,32].

The IS method works by minimizing a simple objective for each variable in the model. The optimal point of this minimization procedure then represents the learned conditional of that variable $[\mu(\sigma_i|\underline{\sigma}_{\setminus i})]$. Given m independent and identically distributed samples (i.i.d.) $\{\underline{\sigma}^{(1)}, \dots, \underline{\sigma}^{(m)}\}$ drawn from a EBM $\mu(\underline{\sigma})$, these representations are learned by minimizing the following loss function for each variable in the model:

$$\begin{aligned} \mathcal{S}_u(\underline{\theta}) &= \frac{1}{m} \sum_{t=1}^m \exp \left\{ -\langle \phi[\sigma_u^{(t)}], \mathbf{f}[\underline{\sigma}_{\setminus u}^{(t)}; \underline{\theta}] \rangle \right\} \\ &= \frac{1}{m} \sum_{t=1}^m \exp \left\{ -\sum_{a=1}^q \phi_a[\sigma_u^{(t)}] f_a[\underline{\sigma}_{\setminus u}^{(t)}; \underline{\theta}] \right\}. \end{aligned} \quad (12)$$

Here, $\phi_a : [q] \rightarrow \mathbb{R}$ is a shifted delta function defined as $\phi_a(x) = \delta_{a,x} - \frac{1}{q}$. Also, $\mathbf{f}(\cdot; \underline{\theta}) : [q]^{n-1} \rightarrow \mathbb{R}^q$ is a vector valued function parameterized by $\underline{\theta}$. For example, $\mathbf{f}(\cdot; \underline{\theta})$ can be a polynomial function from \mathbb{R}^{n-1} to \mathbb{R}^q . In this case, $\underline{\theta}$ would be the set of coefficients that define that polynomial. Another example of $\mathbf{f}(\cdot; \underline{\theta})$ that we use is a neural network. In that case, $\underline{\theta}$ is the set of all trainable parameters (weights and biases)

in the network. The neural network variant of this method is known as *NeuRISE* [29].

The optimum of this loss function is connected to the true model $\mu(\underline{\sigma})$ in the following sense: let $\underline{\theta}^* \equiv \text{argmin } \mathcal{S}_u(\underline{\theta})$. Then we can show that

$$\mu(\sigma_u|\underline{\sigma}_{\setminus u}) \approx \frac{\exp[\langle \phi(\sigma_u), \mathbf{f}(\underline{\sigma}_{\setminus u}; \underline{\theta}^*) \rangle]}{\sum_{a=1}^q \exp[\langle \phi(a), \mathbf{f}(\underline{\sigma}_{\setminus u}; \underline{\theta}^*) \rangle]}. \quad (13)$$

If the parametric function family is expressive enough, then in the $m \rightarrow \infty$ limit, this becomes an exact equality [29].

The value of m required in Eq. (12) to learn a polynomial energy function, up to a certain error, can be bounded using techniques from stochastic convex optimization. The complete analysis can be found in Ref. [28], and the sample complexity of this method is found to asymptotically match known lower bounds [36].

1. Effective temperature and sample complexity

No generative model can learn every quantum state efficiently. For many neural-net-based approaches, there is no theoretical understanding of which states can be hard for a given method. The method we present here is different, as from extensive work on the IS method [28,30,31] and from information-theoretic lower bounds [36], there is an exact characterization of which distributions are hard to learn. These results are best explained by considering the EBM defined in Eq. (2) as a Gibbs distribution of a classical Hamiltonian on n classical spins with $O(1)$ couplings and at an inverse temperature of β . It is known that the number of samples required to learn the energy function in the monomial basis scales as $e^{O(\beta d)} \ln n$, where d is the maximum number of interaction terms (monomials) that are connected to any one variable. This expression tells us that learning is easy for models at a fixed inverse temperature $\beta = O(1)$ defined locally on some lattice of fixed dimension $d = O(1)$.

Physically, the relationship between the sample complexity and inverse temperature can be understood as follows. Suppose we are observing samples being produced from a classical system in thermal equilibrium. At low temperatures, the samples we observe will predominantly be the ground states of this classical system. The amount of samples one needs to draw to observe some excited states above the ground states scales as $\exp[O(\beta d)]$. Since the ground states are themselves not sufficient to uniquely fix the energy function, it is necessary for the user to observe $\exp[O(\beta d)]$ independent samples to see some nontrivial configurations that can be exploited for learning.

As a consequence of the information theoretic lower bounds, we can also deduce which quantum states/POVM combinations are easy or hard to learn from the value of the effective inverse temperature of their EBM representations. The measurement data from a state that correspond to lower absolute value of the coefficients in the energy function (when represented as a polynomial) will be easier to learn than a state that has a higher absolute value of the coefficients in the energy function. This is because the EBM representation of the former quantum state corresponds to a classical Gibbs distribution of lower β . From the theory outlined in Ref. [28], the precise definition of effective inverse temperature of an

EBM that controls the sample complexity can be derived from the L_1 norm of the true parameters of the EBM. However, for the purposes of this paper, we will take the absolute values of the largest coefficients in the model as a measure of inverse temperature. For energy functions that are local, as we expect from physical systems, the notion of temperature used here is essentially the same as the definition used in the rigorous machine-learning literature [28,30,31,43].

2. Function families and neural network architecture

We use three different parametric function families in the IS loss function in this paper. The first is a polynomial *Ansatz*. In this case, the coefficients of the polynomial act as the trainable parameters in the model. This leads to a convex objective function in Eq. (12).

We also use neural nets as a parametric function family in Eq. (12). Any neural network architecture that allows for fast computation of its gradients can be used here. We use a simple feed-forward architecture with fully connected layers for its universal approximating properties and simplicity in implementation. We use Swish as the nonlinear activation in these models [44]. Swish is a continuous variant of the more popular ReLU activation.

For learning permutation invariant states, we use a general symmetric function as a parametric *Ansatz*. For an n -qubit system, this leads to an optimization problem of size $O(qn^{q-1})$. Moreover, we can choose a parameterization of these functions such that the objective in Eq. (12) is a convex function. For more technical details pertaining to these three function families, see Appendix C. The exact sizes and training parameters used for each experiment in this paper are reported in Appendix D.

3. Selection of the parametric family

The EBM learning procedure outlined in the previous sections requires us to choose a parametric function family that can effectively capture the distribution generated by a quantum state/POVM combination. We find that this choice can be effectively made for a family of states by learning a polynomial representation for the energy function for a small state. The polynomial family can be chosen to include higher-order terms which would be computationally infeasible to consider for larger states, as including all terms up to order k would imply a computational cost of $O(n^k)$. Also, for smaller states, it is possible to do learning in the limit of an infinite number of samples drawn from the state ($m \rightarrow \infty$ limit) as the classical distribution generated by a POVM can be computed exactly for smaller states. Learning the energy function in the polynomial basis for smaller states can give us valuable information about the types of terms present in the energy function. This information can then be used to choose the appropriate function family to use for larger states. If there are no obvious symmetries or if there are terms of higher order present, then a neural net family can be chosen to represent the energy function. If energy functions learned from smaller states only show the presence of terms of quadratic or lower order, then a polynomial *Ansatz* would be suitable. If we have prior information about the symmetries of the quantum states,

then these symmetries can be incorporated to reduce the size of the parameter space we must optimize over.

4. Error metrics

The IS method is nearly sample optimal for learning the parameter values in the energy function [31], but for data generated from quantum states, the true parameter values of the energy function are not known. Extracting the parameters in the energy function can be done by brute force for smaller systems but becomes prohibitively expensive even for systems of moderate size. Thus, to compare the error in our learned representations, we compute the error in one- and two-body observables in the L_1 norm, averaged over the entire lattice. For models learned in the computational basis, this reduces to average errors in absolute value in observables of the form $\langle Z \rangle$ and $\langle ZZ \rangle$. For models learned using samples generated by the tetrahedral POVM, we use the trace distance between one- and two-body reduced density matrices as our error metric. This quantity then upper bounds the average error in any one- and two-body observable, up to a constant factor [38].

For experiments involving GHZ states and other permutation-invariant states, we also use estimated fidelity and total variation distance (TVD) as error metrics.

Given a pure state $|\psi\rangle$ and a mixed state ρ , we define the quantum fidelity between these states as the following overlap [45]:

$$F^Q(|\psi\rangle, \rho) \equiv \langle \psi | \rho | \psi \rangle. \quad (14)$$

Expanding ρ in the dual POVM elements as in Eq. (8), we get

$$F^Q(|\psi\rangle, \rho) = \sum_{\underline{\sigma}} \mu(\underline{\sigma}) \langle \psi | D_{\underline{\sigma}} | \psi \rangle. \quad (15)$$

This quantity can be easily estimated from samples drawn from μ , if the state $|\psi\rangle$ has an efficient classical representation, like for the GHZ state.

We also use the classical measure, TVD to compare two probability distributions. The TVD between two distributions ν and μ is simply defined as

$$\text{TVD}(\nu, \mu) = \frac{1}{2} \sum_{\underline{\sigma}} |\nu(\underline{\sigma}) - \mu(\underline{\sigma})|. \quad (16)$$

In our experiments, the distribution μ will be the exact distribution representing a quantum state, as defined in Eq. (1). In case ν represents the learned EBM, the TVD can be estimated by replacing ν with the empirical probabilities obtained from the samples drawn from the EBM.

III. RESULTS

A. Learning mixed states

We start by exploring learning of energy-based representations for the mixed states. Specifically, we focus on the thermal states of the form:

$$\rho = \frac{\exp(-\beta H)}{\text{Tr}[\exp(-\beta H)]}, \quad (17)$$

where H is a local Hamiltonian. For the illustrations in this section, we use the transverse field Ising model (TIM)

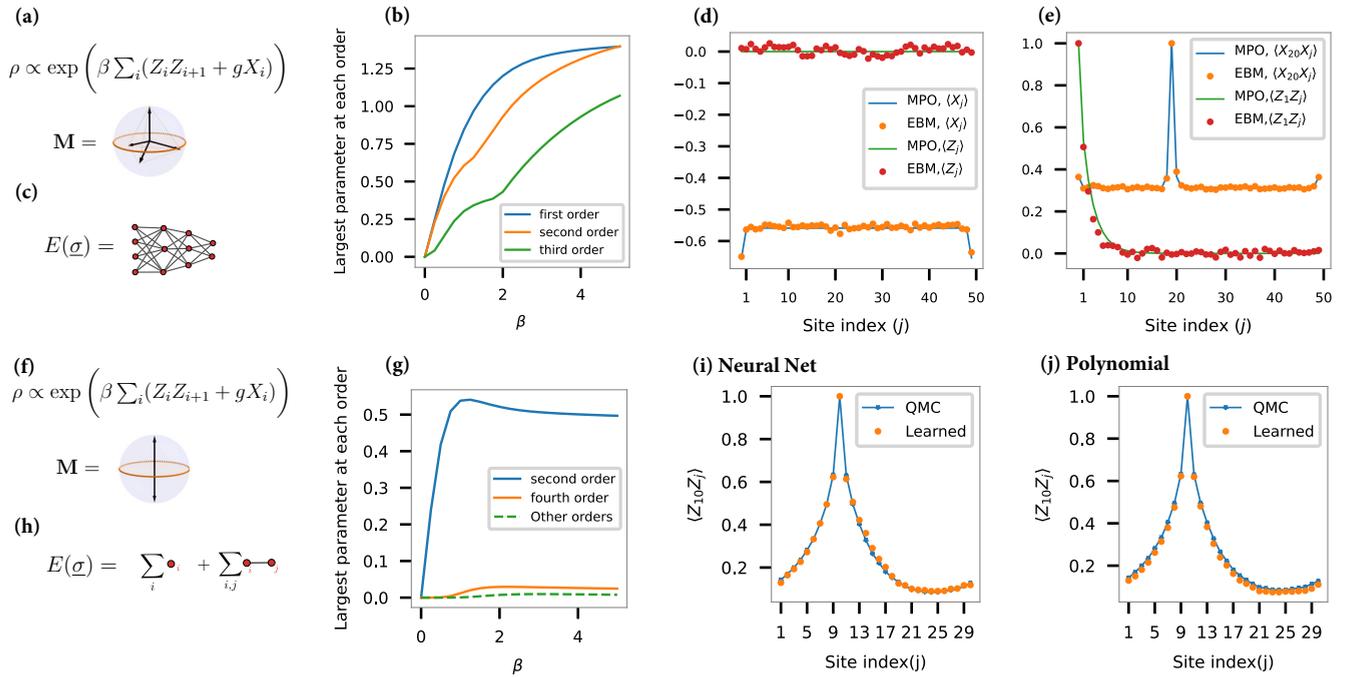


FIG. 3. Learning classical energy-based representations for quantum thermal states. (a) Energy-based model (EBM) learned from the transverse field Ising model (TIM) thermal state measured using the tetrahedral positive operator valued measures (POVM). (b) The strength of interactions at each order for a third-order polynomial energy function exactly representing the thermal state of a 5 qubit TIM on a line measured using the tetrahedral POVM. Here, the absolute value of the largest coefficient in the polynomial representation at each order is plotted against the inverse temperature. For these experiments, we fix $g = 1$. We see that terms at all orders in the energy function are not negligible. (c) A neural net representation is most appropriate, given the presence of higher-order terms in the polynomial representation. (d) One-body and (e) two-body expectation values inferred from the neural net representation for the ferromagnetic one-dimensional (1D) TIM at $\beta = 1$, $g = 1$, with open boundary conditions. The EBM representation is learned using $m = 10^5$ samples, and observables are computed by drawing 4×10^5 samples from the learned model. (f) TIM thermal state measured in the computational basis only. (g) Strength of interactions at each order for an EBM exactly representing the thermal state of a 5 qubit ferromagnetic TIM using measurements in the computational basis. We see that the second-order terms dominate over other terms. (h) This means that the energy function can be well approximated using a second-order polynomial. ZZ expectations estimated from EBMs represented using (i) neural nets and (j) quadratic polynomials, respectively. Both representations are learned for a 30 spin thermal TIM at $\beta = 2$, $g = 1$, with $m = 10^6$ samples, with periodic boundary conditions. The neural net model has 28 380 trainable parameters, while the polynomial model has only 900. Further details about the numerical experiments are given in Appendix D.

family of Hamiltonians with a uniform transverse field, $H_{\text{TIM}} = \sum_{i < j} J_{ij} Z_i Z_j + g \sum_i X_i$. These models are particularly useful for studying properties of the learning algorithms, as they allow for easy generation of samples, either using MPS in one dimension (1D) [6,46] or using quantum Monte Carlo (QMC) methods in higher dimensions owing to the absence of the sign problem in these models [14].

First, we consider the TIM Hamiltonian on a 1D lattice with $J_{i,i+1} = -1 \forall i$ and zero otherwise, measured in the tetrahedral POVM, see Fig. 3(a). We can obtain insight into whether a more specific polynomial function family (rather than the generic neural net one) can effectively capture the distribution generated by a quantum state and POVM combination using the following approach: for a small state, *exact* learning of the energy function in the polynomial basis would give us valuable information about the types of terms present in the energy function and their symmetry properties, as explained in the Methods section. In Fig. 3(b), we generate the exact distribution defined by (1) for a 5 qubit thermal state using matrix exponentiation and learn an EBM for it using the polynomial family with up to third-order terms. We see

that the polynomial representation learned here has significant terms at all orders including order three, which is in contrast with the respective quantum Hamiltonian which only had second-order interactions. This experiment suggests that using the polynomial representation is not very efficient for this state/POVM combination, as the computational complexity of the learning procedure will scale as $\Theta(n^L)$ for a n -qubit state and for higher-order terms of order L . This motivates the use of a fully connected neural net as the parametric function family for the energy function representation.

The specific algorithm in the IS suite of methods adapted to the neural net parametric family is known as NeuRISE (see Appendix B for more details), which has been shown to be able to learn energy functions with higher-order terms in an implicit sense with fewer parameters when the complexity of learning using polynomials scales unfavorably due to the presence of higher-order terms [29]. In Figs. 3(c) and 3(d), we study the ferromagnetic 1D TIM with 50 qubits in the quantum critical phase [47]. We compare the one- and two-body correlations inferred from the learned classical representation to the ones generated from a MPO representation of the state

found by imaginary time evolution using the ITensor library [46], in Figs. 3(c) and 3(d), respectively. We see that the EBM representation can reproduce expectation values of both diagonal and off-diagonal observables.

The nature of the EBM representation can change considerably depending on the POVM used. We demonstrate this using the same thermal ferromagnetic TIM state, now measured in the computational basis rather than using the tetrahedral POVM, see Fig. 3(e). Note that, for the general states, measurements in the computational basis do not provide an informationally complete set sufficient for unambiguously specifying the entire state. Figure 3(f) shows results for the exact learning on a thermal state of a thermal ferromagnetic TIM state from measurements in the computational basis. The results are markedly different from Fig. 3(b): We see that the second-order interactions dominate the energy representation, and hence, the energy function can be well approximated by a quadratic polynomial. In Figs. 3(g) and 3(h), we show that, like the case of the tetrahedral POVM, the correlation functions are well predicted by both general neural net and specialized polynomial parametric representations of the energy function in the computational basis. However, the polynomial representation uses significantly fewer parameters and is ideal for use in a resource-limited setting.

Results showing scaling of error with the size of the system and number of samples for the TIM are given in Fig. 4 and Appendix E. Specifically, Figs. 4(c) and 4(d) show the average error in ZZ expectation values for the two different values of g , with $\beta = 1$ for a two-dimensional (2D) TIM. Here, it is observed that the learning algorithm performs better at higher values of g . This behavior is expected, as for $g \rightarrow \infty$, the learned distribution gets closer to a high-temperature Gibbs distribution in the computational basis, and it is known from prior works that learning high-temperature EBMs is easy [28,30,31]. On the other hand, for $g \rightarrow 0$, we are close to the $\beta = 1$ thermal state of a classical Ising model. Learning lower-temperature classical Gibbs distributions is known to be difficult in general from sample-complexity lower bounds [36]. The results for a fixed β suggest that the value of g plays an important role in determining the final effective temperature of the distribution we are learning, with larger values of g producing classical distributions with a higher effective temperature. In Appendix E, we also observe that the error does not increase considerably with the number of qubits in the system. This is consistent with the logarithmic scaling of sample complexity with the number of variables observed for learning EBMs [28,30,31].

B. Learning ground states

Learning a distribution of classical ground states is hard due to the high sample complexity of methods at lower temperatures. However, the classical distribution of ground states of quantum systems modeled as an EBM has an effective temperature depending on the strength of off-diagonal terms in the Hamiltonian. Our results below show that EBMs learned from the zero-temperature states of quantum Hamiltonians exhibit a certain effective temperature property that can aid the learning process. As an illustration, we consider the problem of learning an informationally complete representation of

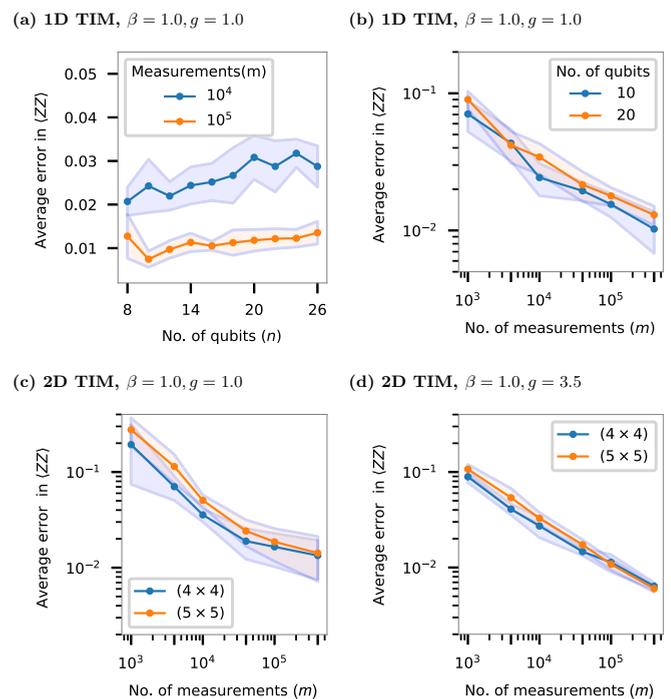


FIG. 4. Scaling results for the classical representations of thermal states of the transverse field Ising model (TIM). Error in ZZ expectation of the polynomial representation for the one-dimensional (1D) TIM at $\beta = 1$: (a) number of qubits (b) number of measurements (i.e., training samples). We see that increasing the number of measurements has decreased the error significantly. On the other hand, the error does not increase appreciably if the number of qubits in the system is increased. (c) Error in ZZ expectation of the polynomial representation for the two-dimensional (2D) TIM at $\beta = 1$, with $g = 1$. (d) The same but for $g = 3.5$. The chosen g values are on either side of the zero-temperature critical point ($g_c = 3.044$) [47]. We see that the model is easier to learn for the larger value of g .

the ground state of an antiferromagnetic TIM from samples obtained from measurements in the tetrahedral POVM, see Fig. 5(a).

As the TIM Hamiltonian is stoquastic, the ground state can be fully specified by computational basis measurements [48,49]. Figure 3(f) shows that, in the regime $\beta \rightarrow \infty$, a polynomial representation for the ground state of the TIM learned from measurements in the Z direction alone will have predominantly second-order terms. However, in general, without prior information about the Hamiltonian, it is hard to determine whether a ground state can be fully specified by only computational basis measurements. Thus, in these experiments, we deliberately do not use the fact that the TIM Hamiltonian is stoquastic and attempt to learn a more general representation of the ground state using the tetrahedral POVM.

In Fig. 5(b), we use our strategy of learning an exact distribution for a small system of size $n = 6$ to get insight into the structure of the energy function. Figure 5(b) shows the strength of terms present in an energy function learned using a third-order polynomial *Ansatz*. Notice that the effective temperature of the EBM is inversely proportional to the strength of interactions shown in Fig. 5(b). We see that, in the small g regime, the strength of interactions blows up as

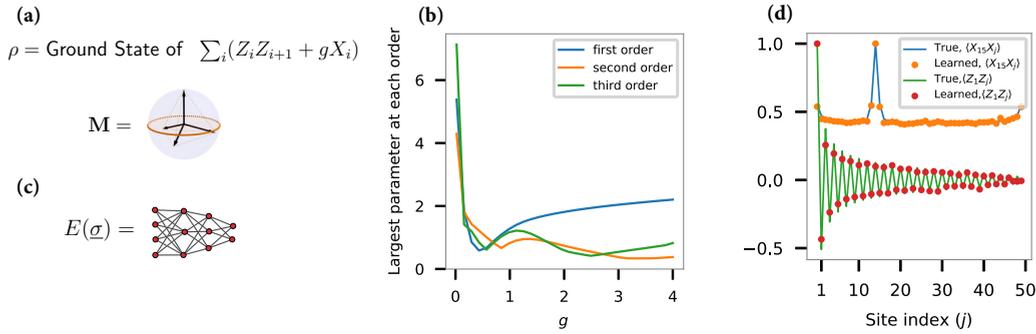


FIG. 5. Learning classical energy-based representations for ground states of quantum systems. (a) Here, we study energy-based model (EBM) representation of a ground state of the antiferromagnetic transverse field Ising model (TIM) measured in the tetrahedral basis. (b) Strength of interactions at each order for an EBM exactly learned from the ground state of a 6 qubit antiferromagnetic TIM on a line. The figure shows that a second-order representation cannot capture all the features of this model. The relative importance of all orders is especially pronounced when the Ising term dominates the transverse field. (c) Because of these higher-order terms, we find that a neural net *Ansatz* is necessary to capture all the features of the model. (d) Two-body expectation values for a 50 spin model at $g = 1$. The EBM learned from this state can accurately learn the long-range behavior of the ground state at the critical point. All results from a neural net *Ansatz* are trained from samples drawn from a matrix product states (MPS) representation of the ground state, found using the density matrix renormalization group (DMRG) [6,46]. The EBM was learned using $m = 10^5$ training samples. Further details about the numerical experiments are given in Appendix D.

$g \rightarrow 0$, thus providing evidence for the low effective temperature of $\mu(\sigma)$ at low values of g . In the large g regime, on the contrary, the first-order terms are dominant, and the effective temperature is not very large. This means that a polynomial model will be able to give a satisfactory approximation in the high g regime. We see that, below the critical point ($g = 1$), terms at all orders are equally important. This means that we will have to resort to a universal neural net *Ansatz* due to the presence of higher-order terms in the energy function.

The effect of g on learning a 60 qubit state from the same family is given in Fig. 6(b) and Appendix E. In line with the results for thermal states in Fig. 4, we see that an EBM can easily be learned for states that possess paramagnetic order. However, as expected, the sample complexity of the learning procedure increases when the states come close to the ground state of the classical model. Therefore, the observed effect of g on learning is much more pronounced for ground states when compared with thermal states.

The results of using NeurISE for an antiferromagnetic TIM on a 50 qubit 1D lattice with open boundaries at the critical point are given in Fig. 5(c). We see that this method has no problems learning states that possess long-range order. This observation is in line with what has been established in classical models as well [31].

Like learning representations of thermal states, our experiments with ground states also show that the error in the learning procedure scales favorably with the system size. In Appendix E, we give scaling results for learning ground states and discuss learning EBM representations of ground states for the Heisenberg model in 2D.

C. Learning states with symmetries

Another important aspect of quantum states that can greatly aid the learning process is the presence of symmetries.

Symmetries can significantly reduce the size of the hypothesis space that a machine-learning method must optimize over. However, it is not always easy to build an *Ansatz* that respects the known symmetries of a model [50–52]. The most common symmetry encountered in physical systems is translational invariance, which can be incorporated into both the energy function representation and the IS learning method rather easily. Once the state is represented by a classical distribution

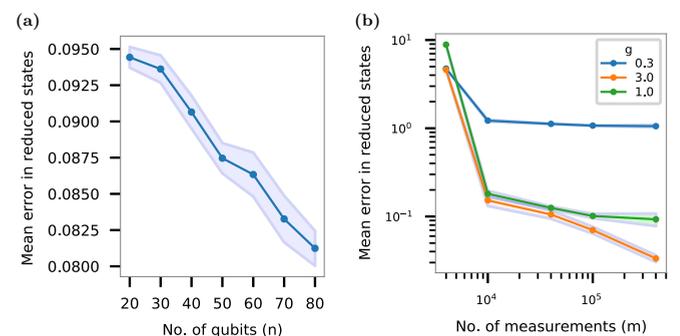


FIG. 6. Scaling results for the classical representations of ground states of the transverse field Ising model (TIM). (a) Error in one- and two-body reduced states as a function of the number of qubits for the ground state of one-dimensional (1D) TIM. The average error goes down in this case, as most errors occur when the spins are close together in the lattice. These energy-based models (EBMs) are learned using $m = 10^5$ samples. (b) For a 60 qubit 1D TIM model, we see the effect of g on learning. The low- g state here is hard to learn, as it is close to the ground state of a classical model. This implies that the effective temperature of the EBM is low, and the sample complexity of learning is high [36]. On the other hand, higher g states are easier to learn, as evidenced by the decreasing error as the number of training samples is increased. Presented is the error (average trace distance between 1 and 2 qubit reduced states) as a function of the number of training samples m .

$\mu(\sigma)$ using the relation in Eq. (1), this distribution inherits the translational symmetry of the state. This in turn implies that the conditional distribution $\mu(\sigma_i|\sigma_{\setminus i})$ of every variable i in the EBM is the same. Since the EBM is learned via these conditionals, the learning step needs only be done for one variable. This leads to a factor n reduction in the computational cost of learning, and nothing extra has to be built into the energy function *Ansatz* that we use. For other types of symmetries, the parametric function family we use must be appropriately chosen to respect the symmetries of the model incorporated into the EBM. This makes learning the representation significantly more efficient due to a reduced number of parameters.

As an example of this approach, we will demonstrate learning states that are invariant under any permutation of qubits. The simplest way to use this prior information is to use a fully symmetric parametric function family to represent the energy function. Any such function acting on n discrete variables where each variable takes q different values can be specified using only $O(n^{q-1})$ parameters. This is because the value of a fully symmetric function can only depend on the number of occurrences of each of the q possible values. For an n -qubit state with the tetrahedral POVM, this implies a representation of size $O(n^3)$, for the most general fully symmetric energy function. We can also use a neural net that respects this permutation symmetry here. However, with a neural net, there is no guarantee that every symmetric function can be represented using a comparable number of parameters. By specifying the most general symmetric function, we can guarantee that the exact EBM that represents the state in question is always considered by the algorithm. Moreover, the optimization problem here is convex in the model parameters compared with the nonconvex optimization required to learn a neural net [29].

To illustrate learning of permutation-symmetric quantum states, we consider learning a classical EBM representation for the GHZ states:

$$|\text{GHZ}_{\pm}^n\rangle = \frac{|0\rangle^{\otimes n} \pm |1\rangle^{\otimes n}}{\sqrt{2}}. \quad (18)$$

Using the exact learning procedure, we find that the EBM representation of these states in the tetrahedral basis has terms of all orders. This explains the results in Ref. [17], where a restricted Boltzmann machine with a few hidden units was found incapable of representing this state when n became large. To learn an energy-based representation, we either must resort to a generic neural net parameterization or incorporate knowledge about the state symmetries. The comparison of these two approaches is given in Fig. 2(a), where we estimate the TVD between the distribution representing $|\text{GHZ}_{+}^7\rangle$ and learned models by drawing new samples from these models. We see that the symmetric *Ansatz* outperforms a neural net *Ansatz* of comparable size. After drawing 4×10^6 samples from these models learned from $m = 10^6$ measurements, we find that the estimated TVD between the symmetric function *Ansatz* and the true state is only 4% more than the inherent finite-sampling error in the estimation procedure, while the TVD estimated for the neural net is 77% more than the sampling error.

In an experiment involving symmetric states, studied in Ref. [15], the fidelity of $|\text{GHZ}_{+}^n\rangle$ is estimated with respect to another mixed state of the form $\rho_{n,p} = (1 -$

$p)|\text{GHZ}_{+}^n\rangle\langle\text{GHZ}_{+}^n| + p|\text{GHZ}_{-}^n\rangle\langle\text{GHZ}_{-}^n|$ for $p \in [0, 1]$. It was been reported that, for this mixed state example, generative models learned using the neural net quantum states (NNQS) *Ansatz* [15,17] struggled to find the correct fidelity as p got closer to 1. The results of applying our method on this problem with a symmetric function *Ansatz* are given in Fig. 2(b). Compared with the NNQS results reported in Ref. [15], we see that the IS method with the symmetric function *Ansatz* can better predict the overlap of the above state with $|\text{GHZ}_{+}^n\rangle$.

IV. DISCUSSION

We have presented an approach for learning generative models for quantum states. The classical energy-based representation used here is flexible enough to incorporate prior information about the state including its symmetries. Using a few representative cases of pure or mixed quantum states, we have illustrated the respective advantages of using different parametric families for modeling energy functions of the emerging classical representations. We have also observed that the classical energy function inherits a certain effective temperature from the quantum states. This effective temperature acts as an information-theoretic metric for the hardness of learning of a classical representation for a given class of quantum states. In this paper, we also demonstrate interesting connections between classical thermal states and quantum states which might be of independent interest to physicists.

There are many interesting directions to explore using the classical energy-based representations of quantum states. An interesting theoretical question is whether the classical energy function inherits locality or low-degree structure from the quantum Hamiltonian. This possibility cannot be ruled out completely from the exact learning experiments presented in this paper. If such a connection can be shown for the EBM for some class of quantum states, then that implies an efficiently learnable representation for those states. Polynomial representations of energy functions are well suited for such a theoretical analysis when compared with black-box techniques that use neural nets. Another important direction is learning from noise-corrupted data, as most real-world tomographic data are corrupted by errors in the measurement process [53,54]. One of the advantages of the IS-based learning methods is that they can be made robust to many types of noisy channels [55]. We leave the development of energy-based learning methods with robustness to measurement noise as future work.

Data used for the plots in this work are available from the corresponding author upon reasonable request. Code to reproduce the experiments described in this paper can be found online [56].

ACKNOWLEDGMENTS

The authors acknowledge support from the Laboratory Directed Research and Development program of Los Alamos National Laboratory (LANL) under Projects No. 20220545CR-NLS, No. 20210674ECR, and No. 20230338ER, and from the U.S. Department of Energy/Office

of Science Advanced Scientific Computing Research Program. This paper used computational resources provided by the Darwin testbed and the Institutional Computing program at LANL.

APPENDIX A: NOTATIONS

Vector-valued functions are denoted by boldface (e.g., \mathbf{f} , Φ). Boldface is removed while referring to their individual outputs (e.g., f_a , Φ_a). Vector variables are denoted by an underline ($\underline{\sigma}$). The underline is not used while referring to their individual outputs (σ_u). The subset of variables from this vector given by an index set K is denoted by $\underline{\sigma}_K$. The shorthand $\underline{\sigma}_{\setminus u}$ is used to denote all the elements of $\underline{\sigma}$ excluding σ_u . Here, $\{X, Y, Z\}$ are reserved for single-qubit Pauli operators.

APPENDIX B: LOCAL ENERGY AND CONDITIONALS OF EBMs

Given a set of n random variables $\underline{\sigma} = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$, an EBM is a positive joint probability distribution over these variables given by

$$\mu(\underline{\sigma}) = \frac{\exp[E(\underline{\sigma})]}{Z}. \quad (\text{B1})$$

Here, E is a real-valued function known as the energy function, and Z is the partition function that ensures normalization. The random variables can be continuous or discrete. In this paper, we focus on the discrete EBMs. Moreover, we assume that each random variable in the model takes values from the set $[q]$. We refer to q as the alphabet size of the model. For each variable σ_u in the model, it is possible to split the energy function into two parts such that one of the parts completely capture the dependence of the energy function on the values taken by that variable alone:

$$\begin{aligned} E(\underline{\sigma}) &= \langle \phi(\sigma_u), \mathbf{E}^u(\underline{\sigma}_{\setminus u}) \rangle + E^u(\underline{\sigma}_{\setminus u}) \\ &= \sum_{a=1}^q \phi_a(\sigma_u) [E_a^u(\underline{\sigma}_{\setminus u}; \underline{\theta})] + E^u(\underline{\sigma}_{\setminus u}). \end{aligned} \quad (\text{B2})$$

Here, $\phi_a(x) = \delta_{a,x} - 1/q$ are centered delta functions, and $\mathbf{E}^u: [q]^{n-1} \rightarrow \mathbb{R}^q$ is a vector-valued function. Together, $\langle \phi(\sigma_u), \mathbf{E}^u(\underline{\sigma}_{\setminus u}) \rangle$ includes in it all the terms of E that are dependent on u . We call this term the *local energy function of u* . The second term $E^u(\underline{\sigma}_{\setminus u})$ includes the rest of the terms that do not depend on the value taken by σ_u . Notice that this decomposition does not put any restriction on our EBM, as any multivariate function with discrete inputs can be decomposed in this way.

The local energy of a variable σ_u is related to its conditional distribution, given all the other variables:

$$\mu[\sigma_u | \underline{\sigma}_{\setminus u}] = \frac{\exp[\langle \phi(\sigma_u), \mathbf{E}^u(\underline{\sigma}_{\setminus u}) \rangle]}{\sum_{s=1}^q \exp[\langle \phi(s), \mathbf{E}^u(\underline{\sigma}_{\setminus u}) \rangle]}. \quad (\text{B3})$$

This makes it possible to sample from the EBM using Gibbs sampling once local energies of every variable are known.

APPENDIX C: PARAMETRIC FUNCTION FAMILIES

Here, we discuss the parametric function families that we use for modeling the EBM energy functions in this paper.

1. Polynomials

Polynomials are the most widely used function family for representing EBMs. For discrete random variables, any energy function can be expressed as a linear combination of a finite number of monomial terms. For the monomial family, the minimization of the IS objective is a convex optimization problem [28,30,31]. The IS method is also known to be sample optimal in this case. We train the polynomial models using the entropic descent algorithm [28,57].

If the true energy function is a polynomial function of degree L , then the complexity of learning a polynomial representation scales as $O(n^L)$. Thus, learning polynomial representations is tractable only if the true energy is given by a low-degree polynomial. Many models used in classical statistical physics have a low degree and can be efficiently learned using polynomials. However, this low-degree condition is not satisfied in general for quantum tomography data.

2. Neural nets

The use of neural nets in the IS method was introduced in Ref. [29]. The neural network *Ansatz* is useful if the EBM being learned has higher-order terms and if there are no obvious symmetries in the problem that can reduce the size of the polynomial *Ansatz*. The IS loss function for this case simply reads

$$\mathcal{S}_u(\underline{\theta}) = \frac{1}{m} \sum_{t=1}^m \exp \left\{ -\langle \phi[\sigma_u^{(t)}], \text{NN}[\underline{\sigma}_{\setminus u}^{(t)}; \underline{\theta}] \rangle \right\}. \quad (\text{C1})$$

Here, we have used a vector-valued neural net for the general function in Eq. (12). The $\underline{\theta}$ parameters are now the weights and biases of the neural net. Since neural nets can approximate any arbitrary function, one can show that this *Ansatz* is powerful enough to learn any energy function.

The minimization of this loss function can be done using a variant of the stochastic gradient descent algorithm. The gradients can be computed using backpropagation. This enables the training of the neural network *Ansatz* on a GPU.

We use feed-forward neural nets as our neural net *Ansatz*. In general, such a net is a function $f: \mathbb{R}^{n-1} \rightarrow \mathbb{R}^q$, defined as

$$\mathbf{f}(\underline{x}) = \mathbf{A}_{d+1} \cdot \alpha(\dots \{ \mathbf{A}_2 \cdot [\alpha(\mathbf{A}_1 \underline{x} + \underline{b}_1)] + \underline{b}_2 \} \dots) + \underline{b}_{d+1}. \quad (\text{C2})$$

Here, $\mathbf{A} \cdot \underline{x} + \underline{b}$ is an affine transformation, and $\alpha: \mathbb{R} \rightarrow \mathbb{R}$, known as the activation function, acts on a vector component wise, i.e., $\alpha(\underline{x}) \equiv [\alpha(x_1), \alpha(x_2), \dots]$. The matrices $\mathbf{A}_1, \dots, \mathbf{A}_{d+1}$ and the vectors $\underline{b}_1, \dots, \underline{b}_{d+1}$ are the trainable parameters of this model.

We call the number d that fixes the number of layers of the net the *depth* of the network. To simplify the architecture of the nets, we force the intermediate layers to all have the same size, i.e., $\mathbf{A}_1 \in \mathbb{R}^{n-1 \times w}$, $\mathbf{A}_2 \dots \mathbf{A}_d \in \mathbb{R}^{w \times w}$, and $\mathbf{A}_{d+1} \in \mathbb{R}^{w \times d}$. We call w the *width* of the net. In general, we can use

a different activation function at each layer, but in this paper, we use this simpler model of a neural net described above.

Through out this paper, we use the Swish function as our activation function [44], $\alpha(x) = \frac{x}{1+\exp(-x)}$. This is a differentiable variant of the popular ReLU activation function used extensively in deep learning. We train our model using a popular variant of the stochastic gradient descent algorithm known as ADAM [58].

3. Symmetric functions

Symmetric functions can be used as *Ansätze* for the local energies to learn a fully symmetric EBM. If we have prior information that the data we use for learning come from a fully symmetric distribution, then using such an *Ansatz* can reduce the size of the model being learned from $O(q^n)$ to $O(qn^{q-1})$. This reduction in size makes the learning problem tractable.

A fully symmetric graphical model will be invariant under any permutation of its variables:

$$\mu(\sigma_1, \sigma_2, \dots, \sigma_n) = \mu[\sigma_{\pi(1)}, \sigma_{\pi(2)}, \dots, \sigma_{\pi(p)}], \quad \forall \pi \in S_n. \quad (C3)$$

Here, S_n is the symmetric group on n elements.

It is straightforward to see that all the single variable conditionals of this distribution will inherit this permutation symmetry:

$$\mu(\sigma_u | \underline{\sigma}_{\setminus u}) = \mu(\sigma_u | \underline{\sigma}_{\pi(\setminus u)}), \quad \forall \pi \in S_{n-1}. \quad (C4)$$

This implies that the true local energies of the model can be represented by symmetric functions on $n - 1$ variables, and we can use this function family in the IS loss function. The outputs of such functions are agnostic to the order of their inputs, i.e., they depend only the number of occurrences of each alphabet. This property can be used to find compact representations for such functions. Let us define a function that counts the occurrence of each alphabet in an input $\ddagger : [q]^{n-1} \rightarrow \mathbb{R}^q$, such that $\ddagger(\underline{\sigma}_{\setminus u})_a$ gives the number of occurrences of a in $\underline{\sigma}_{\setminus u}$. Using this, we can define the following IS loss function:

$$S(\Theta) = \frac{1}{N} \sum_{t=1}^N \exp(-\langle \phi[\sigma_1^{(t)}], \Theta \{ \ddagger[\underline{\sigma}_{\setminus 1}^{(t)}] \} \rangle), \quad (C5)$$

Here, Θ is a vector-valued function that acts as our local energy *Ansatz*. The possible inputs to Θ are integer strings of length q that sum to $n - 1$. There are only $O(n^{q-1})$ such inputs, and we can use the outputs of Θ at each possible input as the optimization variables in our loss function. Also, using these optimization variables leads to a convex loss function and hence fast optimization in practice. The optimization need only be done for the first variable, as all the local energies will be the same due to the symmetry of the model. We perform

this optimization using interior-point methods provided in the IPOPT package [59]. This is guaranteed to find the global optimum of the IS loss function.

Other symmetries can also be incorporated in a similar way by using a variational *Ansatz* for the energy function that is invariant under the symmetry transformations in question. For the any parametric function family, if the symmetry group is small enough, this can be achieved by simple averaging. For instance, a new function $\tilde{\mathbf{f}}$, defined as $\tilde{\mathbf{f}}(\underline{\sigma}, \theta) \equiv \mathbf{f}(\underline{\sigma}, \theta) + \mathbf{f}(-\underline{\sigma}, \theta)$, is by construction symmetric under spin flip and is useful for learning models with a global \mathbb{Z}_2 symmetry.

It is also worth noting that the IS objective can naturally impose translation invariance with any function family. This is because optimizing the objective in Eq. (12) for any function family reconstructs the local piece of the energy function at each site. If the model is translationally invariant, every local piece of the energy function will be the same (i.e., $\mathbf{E}^u = \mathbf{E}^v \forall u, v \in [n]$), and we only need to optimize for a single site.

APPENDIX D: ADDITIONAL DETAILS ABOUT METHODOLOGY

1. Additional details for numerical experiments

In this section, we provide details on the experiments involving learning of EBM representations of quantum systems:

(1) Figure 3(b): Exact distribution μ was obtained by exactly exponentiating the quantum Hamiltonian. This distribution was used to take the $m \rightarrow \infty$ limit in the IS loss function. In the learning process, a third-order polynomial *Ansatz* was optimized using an interior point method from IPOPT.

(2) Figure 3(g): This experiment uses a neural net with depth $d = 3$ and width $w = 15$. The model was trained with a mini-batch size of 500, with an early stopping criteria with a delta of 10^{-4} for a maximum of 1500 epochs. Training was done using ADAM with initial learning rate of 10^{-2} and an inverse time decay schedule.

(3) Figures 5(c) and 6: The experiments in these figures use neural nets with depth $d = 3$ and width $w = 25$. The model was trained with a mini-batch size of 5000, with an early stopping criteria with a delta of 10^{-8} for a maximum of 500 epochs. Training was done using ADAM with initial learning rate of 8×10^{-3} and an inverse time decay schedule.

(4) Figure 3(f): Exact distribution μ was obtained by exactly exponentiating the quantum Hamiltonian. In the learning process, a polynomial *Ansatz* with terms at all orders was optimized using entropic gradient descent.

(5) Figure 2(a): This experiment uses a neural net with depth $d = 2$ and width $w = 8$. The model was trained with a mini-batch size of 3000, for 200 epochs. Training was done using ADAM with initial learning rate of 10^{-2} and an inverse time decay schedule.

(6) Figure 5(b): Exact distribution μ was computed from an MPS representation of the ground state obtained by the density matrix renormalization group (DMRG). In the learning process, a third-order polynomial *Ansatz* was optimized using an interior point method from IPOPT.

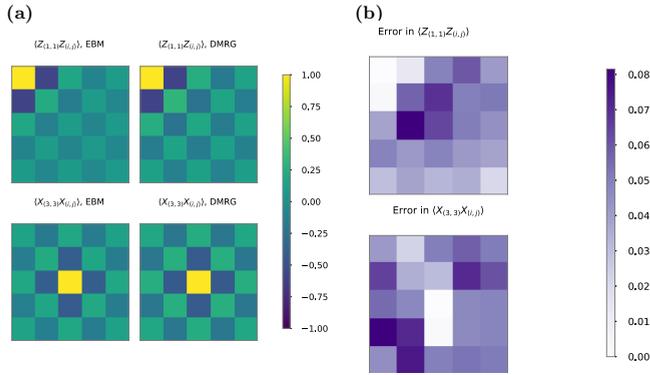


FIG. 7. Learning classical energy-based model (EBM) representation for the ground state of a two-dimensional (2D) Heisenberg model. (a) Comparison of ZZ and XX expectation values learned by the NeuRISE algorithm with values obtained from the density matrix renormalization group (DMRG), $m = 4 \times 10^6$. The Hamiltonian here is the antiferromagnetic Heisenberg model on a 5×5 lattice. (b) Error between expectation values learned by NeuRISE and DMRG. We see good agreement in local observables for learning this 2D model. Just like for the one-dimensional (1D) models discussed in the main text, we also see improved predictions while increasing the training values of m . The neural net used here has only 23 250 free parameters compared with the Hilbert space dimension of $\approx 3 \times 10^7$.

(7) Figure 7: The experiments in these figures use neural nets with depth $d = 3$ and width $w = 15$. The model was trained with a mini-batch size of 10 000, with early stopping criteria with a delta of 10^{-8} for a maximum of 500 epochs. Training was done using ADAM with initial learning rate of 8×10^{-3} and an inverse time decay schedule.

APPENDIX E: ADDITIONAL NUMERICAL RESULTS

1. Scaling experiments for thermal states

Scaling results for learning thermal states using the polynomial *Ansatz* are given in Fig. 4. Figures 4(a) and 4(b) show results for a 1D TIM. We can see that error in the expectation values is most significantly affected by the number of samples used for learning. On the other hand, increasing the number of qubits in the system does not increase the average error in the expectation values by much. Figures 4(c) and 4(d) show similar scaling experiments for a 2D TIM. Here, we see that error has only a weak dependence on the size of the system. The error is seen to decrease consistently with the number of

training samples, but the model with a higher value of g is seen to be easier to learn. This points to an effective temperature in the EBM induced by the transverse field.

2. Additional experiments with ground states

Scaling experiments for learning ground states are given in Fig. 6. Interestingly, in Fig. 6(a), we see that the average error in reduced states decreases with increasing system size. This is due to fact that most errors for 2 qubit reduced states occur when the qubits are closer together on a lattice. For qubits that are farther apart, the EBM reconstructs the reduced states well.

In Fig. 6(b), we study the effect of the phase of the ground state on the learning algorithm for a 1D TIM. Below the critical point of $g = 1$, the ground state is close to the ground state of the classical Ising model. As explained in the Appendices, these states are known to be hard to learn. We see that reflected in these experiments, as the errors in the reduced states do not decrease when we increase the training samples. For higher values of g , the ground state moves away from the classical ground state and this allows for easier learning. This is evidenced by a consistent decrease in error with increasing training samples in these models.

We test the quality of the EBM representation for the ground state of a model other than TIM in Fig. 7, where we study the Heisenberg model in 2D:

$$H = \sum_{\langle i,j \rangle} X_i X_j + Y_i Y_j + Z_i Z_j. \quad (E1)$$

Here, we use a neural net representation for the energy function. The correlation functions obtained from the learned EBM representation show excellent agreement with the ground-truth correlation expectation values.

3. Effect of POVM on learning GHZ state

For learning the GHZ state, we have observed that measurement outcomes from a rotated version of the tetrahedral POVM produce better results than the unrotated version. Table I shows numerical results to this effect. In the experiments involving permutation invariant states, we have chosen one such rotated POVM. The results in Table I suggest that the performance of the algorithm can be improved by choosing an optimal rotation. We leave the exploration of this for future work.

TABLE I. Estimated fidelity of the learned EBM representation with the GHZ state when learned from m measurement shots. Mean and standard errors (in parentheses) are reported from the results of 20 experiments. The rotated tetrahedral POVM is rotated using a different 1 qubit Haar random rotation for every experiment. Reported fidelity values are estimated after drawing 10^6 samples from the learned EBM. We see that the rotated tetrahedral POVM gives a similar error and a much lower variance than the unrotated POVM.

	$m = 5 \times 10^4$	$m = 10^5$	$m = 5 \times 10^5$
Tetrahedral	0.9985 ($\sigma = 0.0385$)	1.0007 ($\sigma = 0.0328$)	0.9988 ($\sigma = 0.0155$)
Rotated tetrahedral	1.0005 ($\sigma = 0.0274$)	1.0040 ($\sigma = 0.0183$)	1.0015 ($\sigma = 0.0081$)

- [1] J. Eisert, D. Hangleiter, N. Walk, I. Roth, D. Markham, R. Parekh, U. Chabaud, and E. Kashefi, Quantum certification and benchmarking, *Nat. Rev. Phys.* **2**, 382 (2020).
- [2] G. Torlai and R. G. Melko, Machine-learning quantum states in the NISQ era, *Annu. Rev. Condens. Matter Phys.* **11**, 325 (2020).
- [3] R. O'Donnell and J. Wright, Efficient quantum tomography, in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, 2016), pp. 899–912.
- [4] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, Sample-optimal tomography of quantum states, in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, 2016), pp. 913–925.
- [5] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Efficient quantum state tomography, *Nat. Commun.* **1**, 149 (2010).
- [6] U. Schollwöck, The density-matrix renormalization group in the age of matrix product states, *Ann. Phys.* **326**, 96 (2011).
- [7] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Quantum state tomography via compressed sensing, *Phys. Rev. Lett.* **105**, 150401 (2010).
- [8] B. P. Lanyon, C. Maier, M. Holzäpfel, T. Baumgratz, C. Hempel, P. Jurcevic, I. Dhand, A. S. Buyskikh, A. J. Daley, M. Cramer *et al.*, Efficient tomography of a quantum many-body system, *Nat. Phys.* **13**, 1158 (2017).
- [9] A. Rocchetto, Stabiliser states are efficiently PAC-learnable, *Quantum Inf. Comput.* **18**, 541 (2018).
- [10] J. Haah, R. Kothari, and E. Tang, Learning quantum Hamiltonians from high-temperature Gibbs states and real-time evolutions, *Nat. Phys.* **20**, 1027 (2024).
- [11] A. Anshu, S. Arunachalam, T. Kuwahara, and M. Soleimanifar, Sample-efficient learning of quantum many-body systems, in *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2020), pp. 685–691.
- [12] C. Rouzé and D. S. França, Learning quantum many-body systems from a few copies, *Quantum* **8**, 1319 (2024).
- [13] E. Onorati, C. Rouzé, D. S. França, and J. D. Watson, Efficient learning of ground and thermal states within phases of matter, [arXiv:2301.12946](https://arxiv.org/abs/2301.12946).
- [14] A. W. Sandvik, An introduction to quantum Monte Carlo methods, in *Strongly Correlated Magnetic and Superconducting Systems*, edited by G. Sierra and M. A. Martín-Delgado (Springer, Berlin, 1997), pp. 109–135.
- [15] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nat. Phys.* **16**, 1050 (2020).
- [16] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, The randomized measurement toolbox, *Nat. Rev. Phys.* **5**, 9 (2022).
- [17] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Reconstructing quantum states with generative models, *Nat. Mach. Intell.* **1**, 155 (2019).
- [18] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Symmetric informationally complete quantum measurements, *J. Math. Phys.* **45**, 2171 (2004).
- [19] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
- [20] G. Torlai and R. G. Melko, Latent space purification via neural density operators, *Phys. Rev. Lett.* **120**, 240503 (2018).
- [21] A. Melkani, C. Gneiting, and F. Nori, Eigenstate extraction with neural-network tomography, *Phys. Rev. A* **102**, 022412 (2020).
- [22] F. Vicentini, R. Rossi, and G. Carleo, Positive-definite parametrization of mixed quantum states with deep neural networks, [arXiv:2206.13488](https://arxiv.org/abs/2206.13488).
- [23] M. Mézard, G. Parisi, N. Sourlas, G. Toulouse, and M. Virasoro, Replica symmetry breaking and the nature of the spin glass phase, *J. Phys. France* **45**, 843 (1984).
- [24] J. Xie, Y. Lu, S.-C. Zhu, and Y. N. Wu, A theory of generative ConvNet, in *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), New York City, NY, USA, June 19–24, 2016*, edited by M.-F. Balcan and K. Q. Weinberger, JMLR Workshop and Conference Proceedings, Vol. 48 (JMLR.org, 2016), pp. 2635–2644.
- [25] T. Kim and Y. Bengio, Deep directed generative models with energy-based probability estimation, [arXiv:1606.03439](https://arxiv.org/abs/1606.03439).
- [26] Y. Du and I. Mordatch, Implicit generation and modeling with energy based models, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, 2019), Vol. 32.
- [27] R. Kumar, S. Ozair, A. Goyal, A. Courville, and Y. Bengio, Maximum entropy generators for energy-based models, [arXiv:1901.08508](https://arxiv.org/abs/1901.08508).
- [28] M. Vuffray, S. Misra, and A. Lokhov, Efficient learning of discrete graphical models, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, 2020), Vol. 33.
- [29] A. J. A. Lokhov, S. Misra, and M. Vuffray, Learning of discrete graphical models with neural networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, 2020), Vol. 33.
- [30] M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov, Interaction screening: Efficient and sample-optimal learning of Ising models, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, 2016), Vol. 29.
- [31] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, Optimal structure and parameter learning of Ising models, *Sci. Adv.* **4**, e1700791 (2018).
- [32] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).
- [33] H. G. Katzgraber, M. Körner, and A. P. Young, Universality in three-dimensional Ising spin glasses: A Monte Carlo study, *Phys. Rev. B* **73**, 224432 (2006).
- [34] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, Distributed message passing for large scale graphical models, in *CVPR 2011* (IEEE, 2011), pp. 1833–1840.
- [35] C.-W. Liu, A. Polkovnikov, A. W. Sandvik, and A. P. Young, Universal dynamic scaling in three-dimensional Ising spin glasses, *Phys. Rev. E* **92**, 022128 (2015).
- [36] N. P. Santhanam and M. J. Wainwright, Information-theoretic limits of selecting binary graphical models in high dimensions, *IEEE Trans. Inf. Theory* **58**, 4117 (2012).
- [37] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, High-dimensional Ising model selection using l_1 -regularized logistic regression, *Ann. Stat.* **38**, 1287 (2010).

- [38] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2002).
- [39] T. Singal, F. B. Maciejewski, and M. Oszmaniec, Implementation of quantum measurements using classical resources and only a single ancillary qubit, *npj Quantum Inf.* **8**, 82 (2022).
- [40] M. Oszmaniec, L. Guerini, P. Wittek, and A. Acín, Simulating positive-operator-valued measures with projective measurements, *Phys. Rev. Lett.* **119**, 190501 (2017).
- [41] C. Chow and C. Liu, Approximating discrete probability distributions with dependence trees, *IEEE Trans. Inf. Theory* **14**, 462 (1968).
- [42] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).
- [43] A. Klivans and R. Meka, Learning graphical models using multiplicative weights, in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (2017), pp. 343–354.
- [44] P. Ramachandran, B. Zoph, and Q. V. Le, Searching for activation functions, [arXiv:1710.05941](https://arxiv.org/abs/1710.05941).
- [45] R. Jozsa, Fidelity for mixed quantum states, *J. Mod. Opt.* **41**, 2315 (1994).
- [46] M. Fishman, S. R. White, and E. M. Stoudenmire, The ITensor software library for tensor network calculations, *SciPost Phys. Codebases* **4** (2022).
- [47] S. Sachdev, *Quantum Phase Transitions*, 2nd ed. (Cambridge University Press, Cambridge, 2011).
- [48] S. Bravyi, D. P. Divincenzo, R. Oliveira, and B. M. Terhal, The complexity of stoquastic local Hamiltonian problems, *Quantum Inf. Comput.* **8**, 361 (2008).
- [49] S. Bravyi, Monte Carlo simulation of stoquastic Hamiltonians, *Quantum Inf. Comput.* **15**, 1122 (2015).
- [50] E. Barnard and D. Casasent, Invariance and neural nets, *IEEE Trans. Neural Netw.* **2**, 498 (1991).
- [51] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, Spherical CNNs, in *6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, April 30–May 3, 2018*, Conference Track Proceedings (OpenReview.net, 2018).
- [52] K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, SchNet: A continuous-filter convolutional neural network for modeling quantum interactions, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, 2017), Vol. 30.
- [53] F. B. Maciejewski, Z. Zimborás, and M. Oszmaniec, Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography, *Quantum* **4**, 257 (2020).
- [54] M. R. Geller and M. Sun, Toward efficient correction of multiqubit measurement errors: Pair correlation method, *Quantum Sci. Technol.* **6**, 025009 (2021).
- [55] S. Goel, D. M. Kane, and A. R. Klivans, Learning Ising models with independent failures, in *Conference on Learning Theory (COLT 2019), Phoenix, AZ, USA, 25–28 June 2019*, edited by A. Beygelzimer and D. Hsu, Proceedings of Machine Learning Research, Vol. 99 (PMLR, 2019), pp. 1449–1469.
- [56] <https://github.com/abhijithjlan1/EBM-Tomography>.
- [57] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Oper. Res. Lett.* **31**, 167 (2003).
- [58] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 7–9, 2015*, edited by Y. Bengio and Y. LeCun, Conference Track Proceedings (OpenReview.net, 2015), [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [59] A. Wächter and L. T. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming, *Math. Program.* **106**, 25 (2006).