

## Recurrent neural network wave functions

Mohamed Hibat-Allah<sup>1,2,3,\*</sup>, Martin Ganahl,<sup>2</sup> Lauren E. Hayward,<sup>2</sup> Roger G. Melko,<sup>2,3</sup> and Juan Carrasquilla<sup>1,3</sup>

<sup>1</sup>Vector Institute, MaRS Centre, Toronto, Ontario, Canada M5G 1M1

<sup>2</sup>Perimeter Institute for Theoretical Physics, 31 Caroline Street North, Waterloo, Ontario, Canada N2L 2Y5

<sup>3</sup>Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1



(Received 9 March 2020; accepted 7 May 2020; published 17 June 2020)

A core technology that has emerged from the artificial intelligence revolution is the recurrent neural network (RNN). Its unique sequence-based architecture provides a tractable likelihood estimate with stable training paradigms, a combination that has precipitated many spectacular advances in natural language processing and neural machine translation. This architecture also makes a good candidate for a variational wave function, where the RNN parameters are tuned to learn the approximate ground state of a quantum Hamiltonian. In this paper, we demonstrate the ability of RNNs to represent several many-body wave functions, optimizing the variational parameters using a stochastic approach. Among other attractive features of these variational wave functions, their autoregressive nature allows for the efficient calculation of physical estimators by providing independent samples. We demonstrate the effectiveness of RNN wave functions by calculating ground-state energies, correlation functions, and entanglement entropies for several quantum spin models of interest to condensed-matter physicists in one and two spatial dimensions.

DOI: [10.1103/PhysRevResearch.2.023358](https://doi.org/10.1103/PhysRevResearch.2.023358)

### I. INTRODUCTION

The last decade has marked the start of a worldwide artificial intelligence (AI) revolution, which is dramatically affecting industry, science, and society. The source of the current AI resurgence can largely be traced back to AlexNet [1], one of the most influential breakthrough papers in computer vision, which provided a dramatic quantitative improvement in object recognition tasks and popularized the paradigm of deep learning [2]. The concept of deep learning encompasses a set of machine learning techniques where data are processed through the composition of parametrized nonlinear layers, each of which generates increasingly abstract representations of the original data [2]. This paradigm has demonstrated an unprecedented unifying power by making advances in areas as diverse as image recognition [3], natural language processing [4], drug discovery [5], self-driving cars [6], game play [7], and more.

The striking performance of deep learning methods has motivated researchers to use a machine learning perspective to reexamine problems in the physical sciences, including areas such as particle physics, cosmology, materials science, quantum chemistry, and statistical physics [8]. The exploration of machine learning techniques has been particularly prominent in the field of quantum many-body physics, where the task of elucidating the equilibrium and nonequilibrium

properties of interacting many-particle systems remains at the research frontier of quantum information and condensed-matter physics. One of the first successful technology transfers from machine learning into many-body physics involved the use of neural network methods in a variational calculation [9]. The variational principle is the theoretical bedrock behind many of the most powerful numerical approaches to solving many-body problems in quantum mechanics [10–12]. Modern incarnations range from well-established techniques such as variational Monte Carlo (VMC) [13] and tensor networks [14] to variational quantum eigensolvers for quantum computation [15]. The resurgence of interest in machine learning has motivated a rich new playground for variational calculations based on a neural network ansatz [9,16–19]. Simultaneous to the computer vision revolution, a wide array of model architectures and algorithmic advances have also emerged in the context of natural language processing (NLP)—the technology that enables computers to process and understand human language. Some of the most important algorithmic advances in NLP have been developed in the context of sequence learning using *recurrent neural networks* (RNNs) [20–24]. These have resulted in impressive results in speech and text comprehension, as well as in state-of-the-art results in neural machine translation. With RNNs and other algorithmic and conceptual advances, algorithms are bringing machine translation and speech recognition closer to the human level with unprecedented success [23,25–27]. Here we explore whether the power and scalability of NLP models such as the RNN can be extended to applications in physical systems, in particular to perform variational calculations to find the low-energy states of quantum many-body Hamiltonians.

RNNs have already proven to be powerful tools within the field of many-body physics. In Ref. [28], RNNs were applied in the context of quantum state tomography and were found to

\*mohamed.hibat.allah@uwaterloo.ca

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.

be capable of representing a broad range of complex quantum systems, including prototypical states in quantum information and ground states of local spin models. Furthermore, RNNs have established similarities to matrix product states (MPS) and are capable of capturing entanglement properties of quantum many-body systems [29]. To date, however, little effort has been made to develop NLP technology for use together with the variational principle. Here we investigate the power of RNNs and their extensions for approximating the ground state of strongly correlated local Hamiltonians. We demonstrate how the variational principle can be combined with RNNs to yield highly efficient ansatz wave functions. Our proposal makes use of the *autoregressive* property [30–32] of RNNs, which, unlike traditional VMC methods, allows for sampling from the wave function. We variationally optimize our RNNs to approximate ground states of various strongly correlated quantum systems in one and two dimensions. We find excellent agreement for local correlation functions and entanglement entropy upon comparison with well-established state-of-the-art approaches, while requiring only a fraction of the variational parameters. Through extensive scaling studies, we show that the intrinsic bias of our ansatz can be systematically reduced to yield highly accurate ground-state approximations of large quantum systems.

## II. CLASSICAL AND QUANTUM RECURRENT NEURAL NETWORKS

### A. RNNs for classical probability distributions

We consider probability distributions defined over a discrete sample space, where a single configuration consists of a list  $\sigma \equiv (\sigma_1, \sigma_2, \dots, \sigma_N)$  of  $N$  variables  $\sigma_n$ , and  $\sigma_n \in \{0, 1, \dots, d_v - 1\}$ . Here, the *input dimension*  $d_v$  represents the number of possible values that any given variable  $\sigma_n$  can take. A central task in machine learning is to use a set of empirical samples to infer probability distributions in cases where there are strong correlations among the variables  $\sigma_n$ . We denote the probability of a configuration  $\sigma$  by  $P(\sigma) \equiv P(\sigma_1, \sigma_2, \dots, \sigma_N)$ , and use the product rule for probabilities to express this distribution as

$$P(\sigma) = P(\sigma_1)P(\sigma_2|\sigma_1) \cdots P(\sigma_N|\sigma_{N-1}, \dots, \sigma_2, \sigma_1), \quad (1)$$

where  $P(\sigma_i|\sigma_{i-1}, \dots, \sigma_2, \sigma_1) \equiv P(\sigma_i|\sigma_{<i})$  is the conditional distribution of  $\sigma_i$  given a configuration of all  $\sigma_j$  with  $j < i$ .

Specifying every conditional probability  $P(\sigma_i|\sigma_{<i})$  gives a full characterization of any possible distribution  $P(\sigma)$ , but in general such a representation grows exponentially with system size  $N$ . Typically, real-world distributions are assumed to endow enough structure on the problem to allow for accurate approximate descriptions of  $P(\sigma)$  that use far fewer resources [33]. This assumption is also applicable in the context of ground-state wave functions that arise in physical systems, which we will discuss at length in this paper.

RNNs form a class of correlated probability distributions of the form Eq. (1), where the  $P(\sigma)$  are entirely specified through the conditionals  $P(\sigma_i|\sigma_{<i})$ . The elementary building block of an RNN is a *recurrent cell*, that has emerged in different versions in the past [24]. In its simplest form, a recurrent cell is a nonlinear function that maps the direct sum (or concatenation) of an incoming *hidden* vector  $\mathbf{h}_{n-1}$  of

dimension  $d_h$  and an input vector  $\sigma_{n-1}$  to an output hidden vector  $\mathbf{h}_n$  of dimension  $d_h$  such that

$$\mathbf{h}_n = f(W[\mathbf{h}_{n-1}; \sigma_{n-1}] + \mathbf{b}), \quad (2)$$

where  $f$  is a nonlinear *activation function*.

The parameters of this simple RNN (“vanilla” RNN) are given by the weight matrix  $W \in \mathbb{R}^{d_h \times (d_h + d_v)}$ , the bias vector  $\mathbf{b} \in \mathbb{R}^{d_h}$ , and the states  $\mathbf{h}_0$  and  $\sigma_0$  that initialize the recursion. In this paper, we fix  $\mathbf{h}_0$  and  $\sigma_0$  to constant values. The vector  $\sigma_n$  is a one-hot encoding of the input  $\sigma_n$  such that, e.g.,  $\sigma_n = (1, 0)$ ,  $(0, 1)$  for  $\sigma_n = 0, 1$  (respectively) when the input dimension is 2. The computation of the full probability  $P(\sigma)$  is carried out by sequentially computing the conditionals, starting with  $P(\sigma_1)$ , as

$$P(\sigma_n|\sigma_{n-1}, \dots, \sigma_1) = \mathbf{y}_n \cdot \sigma_n,$$

where the right-hand side contains the usual scalar product between vectors and

$$\mathbf{y}_n \equiv S(U\mathbf{h}_n + \mathbf{c}). \quad (3)$$

Here,  $U \in \mathbb{R}^{d_v \times d_h}$  and  $\mathbf{c} \in \mathbb{R}^{d_v}$  are weights and biases of a so-called Softmax layer, and the Softmax activation function  $S$  is given by

$$S(v_n) = \frac{\exp(v_n)}{\sum_i \exp(v_i)}.$$

In Eq. (3)  $\mathbf{y}_n = (y_n^1, \dots, y_n^{d_v})$  is a  $d_v$ -component vector of positive, real numbers summing up to 1, i.e.,

$$\|\mathbf{y}_n\|_1 = 1, \quad (4)$$

and thus forms a probability distribution over the states  $\sigma_n$ . Once the vectors  $\mathbf{y}_n$  have been specified, the full probability  $P(\sigma)$  is given by

$$P(\sigma) = \prod_{n=1}^N \mathbf{y}_n \cdot \sigma_n.$$

Note that  $P(\sigma)$  is already properly normalized to unity such that

$$\|P(\sigma)\|_1 = 1. \quad (5)$$

Sampling from an RNN probability distribution is achieved in a similar sequential fashion. To generate a sample  $\sigma = (\sigma_1, \dots, \sigma_N)$  consisting of a set of  $N$  configurations  $\sigma_n$ , one first calculates the hidden state  $\mathbf{h}_1$  and the probability  $\mathbf{y}_1$  from the initial vectors  $\mathbf{h}_0$  and  $\sigma_0$ . A sample  $\sigma_1$  from the probability distribution  $\mathbf{y}_1$  is drawn, which is then fed as a one-hot vector  $\sigma_1$  along with  $\mathbf{h}_1$  back into the recurrent cell to obtain  $\mathbf{y}_2, \mathbf{h}_2$ , and then  $\sigma_2$ . The procedure is then iterated until  $N$  configurations  $\sigma_n$  have been obtained as illustrated in Fig. 1(c).

From Eqs. (2) and (3), it is evident that the hidden vector  $\mathbf{h}_n$  encodes information about previous spin configurations  $\sigma_{<n}$ . For correlated probabilities, the history  $\sigma_{<n}$  is relevant to the prediction of the probabilities of the following  $\sigma_n$ . By passing on hidden states in Eq. (3) between sites, the RNN is capable of modeling strongly correlated distributions. Hereafter, we shall call the dimension  $d_h$  of the hidden state  $\mathbf{h}_n$  the *number of memory units*. We emphasize that the weights  $W$  and  $U$  and

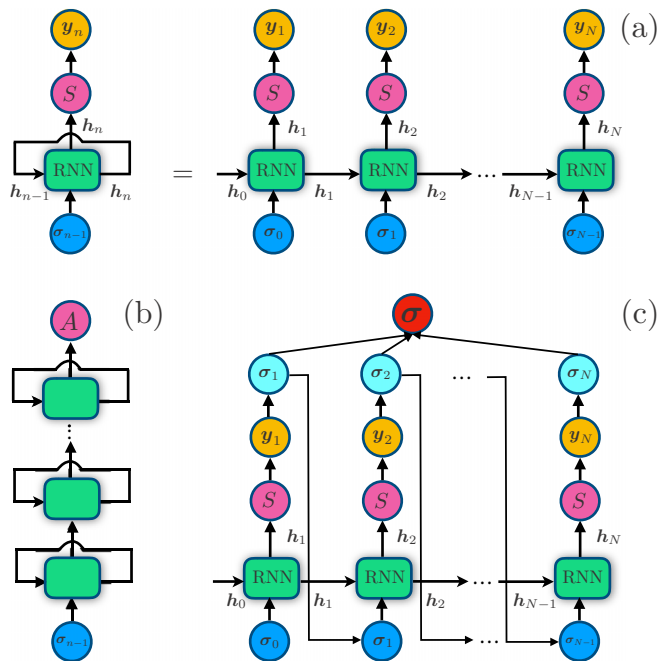


FIG. 1. (a) Left-hand side: An RNN cell (green box) takes a sequence of inputs  $\{\sigma_n\}$ , where at each step  $n$  the input  $\sigma_{n-1}$  and the vector  $h_{n-1}$  are fed in the RNN cell which generates a vector  $h_n$  called the hidden state of the RNN.  $h_n$  is meant to encode the history of the previous inputs  $\sigma_{n' < n}$ . Moreover, the hidden state  $h_n$  is fed to a fully connected layer with Softmax activation  $S$  (magenta circles) to compute conditional probabilities. Right-hand side: The unrolled version of the RNN layer on the left-hand side. (b) A deep RNN model with  $N_l$  stacked single RNN cells (green blocks) followed by a fully connected layer with activation function  $A$  (magenta circle). Each single RNN cell at the  $\ell$ th layer has its corresponding hidden state  $h_n^\ell$ , which serves also as an input for the RNN cell at the  $(\ell + 1)$ th layer. (c) A graphical representation of autoregressive sampling of RNNs.

the biases  $\mathbf{b}$  and  $\mathbf{c}$  together comprise the variational parameters of our ansatz wave function of the next section. These parameters are typically shared among the different values of  $n$ , giving rise to a highly compact parametrization of the probability distribution. Once the dimension  $d_h$  is specified, the number of parameters in the ansatz is independent of the system size  $N$ .

By construction, the model allows for an efficient estimation of the normalized probability of a given configuration  $\sigma$ . This construction is unlike energy-based models, which require intractable calculations of the partition function, or likelihood-free models such as generative adversarial networks that do not allow for an explicit estimation of probabilities [33,34]. The sequential process of computing the probability vectors  $y_n$  is schematically depicted in Fig. 1(a). Deep architectures can be obtained by stacking several RNN cells as shown in Fig. 1(b) for a general activation function  $A$  (not necessarily Softmax). As illustrated in Fig. 1(c), RNNs have the *autoregressive property*, meaning that the conditional probability  $P(\sigma_n | \sigma_{<n})$  depends only on configurations  $\sigma_1, \dots, \sigma_{n-1}$ . We also note that the computational cost of sampling a configuration  $\sigma_1, \dots, \sigma_N$  is linear in the length

$N$  of the configuration. Another important property of the normalized RNN probability distribution is that it can be used to produce successive samples  $\sigma$  and  $\sigma'$  that are independent. Taking advantage of this property, the sampling procedure can be parallelized.

In practice, training vanilla RNNs can be challenging, since capturing long-distance correlations between the variables  $\sigma_n$  tends to make the gradients either explode or vanish [23,35–37]. Similar to MPS [38], long-distance correlations in RNNs are suppressed exponentially [39] and extensions of the vanilla RNN have been proposed [20,40] in order to improve on this limitation. Two successful examples are the long short-term memory (LSTM) unit [20] and the gated recurrent unit (GRU) [40]. Unless stated otherwise, in this paper we use the GRU [40] as the elementary cell in our (one-dimensional) RNNs to study models in one and two spatial dimensions. The details of the implementation can be found in Appendix A.

Furthermore, we explore the use of two-dimensional (2D) vanilla RNNs [21], where information about the spatial location of neighboring spins is exploited by the RNN ansatz. The basic idea of 2D RNNs is to replace the single recurrent connection in a standard RNN, as shown in Eq. (2), with two recurrent connections that are passed to the neighboring sites. Thus, at each point in the lattice the hidden layer of the network receives both spin configuration inputs and the hidden vectors from the neighboring sites, in a way that respects the autoregressive property. We provide the details of the implementation in Sec. III C and Appendix B.

## B. RNN wave functions

The previous section focused exclusively on the efficient parametrization of classical probability distributions  $P(\sigma)$ . In contrast, quantum-mechanical wave functions are in general a set of complex valued amplitudes  $\psi(\sigma)$ , rather than conventional probabilities. Before discussing how to modify the RNN ansatz to represent complex wave functions, we note that an important class of *Stoquastic* many-body Hamiltonians has ground states  $|\Psi\rangle$  with real and positive amplitudes in the standard product spin basis [41]. Thus, these ground states have representations in terms of probability distributions:

$$|\Psi\rangle = \sum_{\sigma} \psi(\sigma) |\sigma\rangle = \sum_{\sigma} \sqrt{P(\sigma)} |\sigma\rangle. \quad (6)$$

This property has been exploited extensively in wave function representations using generative models such as restricted Boltzmann machines [19]. For such wave functions, it is also natural to try to approximate  $P(\sigma)$  with a conventional RNN, as illustrated in Fig. 2(a). For later reference we call this architecture a *positive recurrent neural network wave function* (pRNN wave function).

The generalization to the complex case starts by splitting the wave function into an amplitude and phase  $\phi(\sigma)$  [42] as

$$|\Psi\rangle = \sum_{\sigma} \exp[i\phi(\sigma)] \sqrt{P(\sigma)} |\sigma\rangle. \quad (7)$$

As illustrated in Fig. 2(b), we use one RNN cell and a Softmax layer to model the probability, together with a Softsign layer (as defined below) to model the phase.

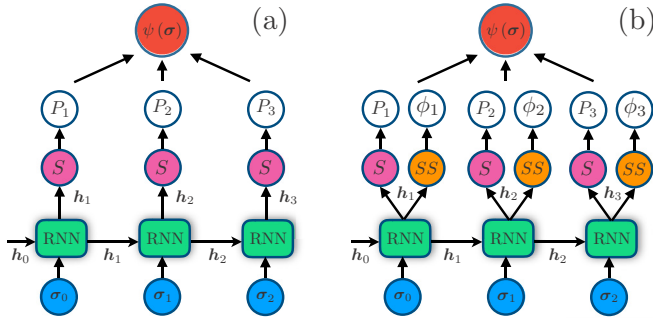


FIG. 2. (a) pRNN wave function: A graphical representation of the computation of positive amplitudes using one RNN cell along with a Softmax layer (magenta circles) to compute the modulus  $|\psi(\sigma)|^2 = P(\sigma)$ . (b) cRNN wave function: A graphical representation of the computation of complex amplitudes using one RNN cell along with a Softmax layer (magenta circles) and a Softsign (SS) layer (orange circles). The first computes the modulus  $|\psi(\sigma)|^2 = P(\sigma)$ ; the second computes the phase  $\phi(\sigma)$  of  $\psi(\sigma)$ .

In this parametrization, the first layer uses the Softmax activation function to get conditional probabilities  $P_n$  as

$$P_n = y_n^{(1)} \cdot \sigma_n, \quad (8)$$

where

$$y_n^{(1)} = S(U^{(1)}\mathbf{h}_n + \mathbf{c}^{(1)}), \quad (9)$$

in a similar fashion to Eq. (3). The Softsign layer is used to compute the phases as

$$\phi_n = y_n^{(2)} \cdot \sigma_n, \quad (10)$$

where

$$y_n^{(2)} = \pi \text{Softsign}(U^{(2)}\mathbf{h}_n + \mathbf{c}^{(2)}). \quad (11)$$

The Softsign function is defined as

$$\text{Softsign}(x) = \frac{x}{1 + |x|} \in (-1, 1).$$

Finally, the probability  $P(\sigma)$  is obtained from the  $N$  individual contributions  $P_n$  as

$$P(\sigma) \equiv \prod_{n=1}^N P_n, \quad (12)$$

and, similarly, the phase  $\phi(\sigma)$  is computed as

$$\phi(\sigma) \equiv \sum_{n=1}^N \phi_n. \quad (13)$$

Note that sampling from the square of the amplitudes  $P(\sigma)$  is unaffected by the Softsign layer and is carried out, as described above, using only the Softmax layer as in Fig. 1(c). For later reference, we call this architecture a *complex recurrent neural network wave function* (cRNN wave function), and, hereafter, the term *RNN wave function* will refer to both pRNN wave functions and cRNN wave functions. Details about the dimensions of the variational parameters of RNN wave functions can be found in Appendix A.

### III. GROUND STATES WITH RNN WAVE FUNCTIONS

We focus our attention on the ground-state properties of prototypical Hamiltonians in condensed-matter physics including the one-dimensional (1D) and 2D transverse field Ising model (TFIM), as well as the 1D  $J_1$ - $J_2$  model, both with open boundary conditions. Their Hamiltonians are given by

$$\hat{H}_{\text{TFIM}} = - \sum_{\langle i,j \rangle} \hat{\sigma}_i^z \hat{\sigma}_j^z - h \sum_i \hat{\sigma}_i^x, \quad (14)$$

where  $\hat{\sigma}_i^{(x,y,z)}$  are Pauli matrices acting on site  $i$ , and

$$\hat{H}_{J_1-J_2} = J_1 \sum_{\langle i,j \rangle} \hat{S}_i \cdot \hat{S}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \hat{S}_i \cdot \hat{S}_j, \quad (15)$$

where  $\hat{S}_i$  is a spin-1/2 operator. Here,  $\langle i, j \rangle$  and  $\langle\langle i, j \rangle\rangle$  denote nearest- and next-nearest-neighbor pairs, respectively. Energies for the  $J_1$ - $J_2$  model are measured in units of  $J_1 = 1$  in the results that follow.

To train our models we use the variational principle, where for a given problem Hamiltonian  $\hat{H}$ , the optimization strategy involves minimizing the expectation value  $E_\lambda = \langle \Psi_\lambda | \hat{H} | \Psi_\lambda \rangle \geq E_0$  with respect to the variational parameters  $\lambda$ . Here,  $E_0$  is the exact ground-state energy of  $\hat{H}$ . The variational parameters  $\lambda$  are updated using variants of the gradient descent algorithm with the objective of minimizing  $E_\lambda = \langle \Psi_\lambda | \hat{H} | \Psi_\lambda \rangle$ . We provide a detailed description of the VMC scheme and the optimization strategy with which we optimize our RNN wave functions in Appendix C.

Since the TFIM in Eq. (14) is Stoquastic, the ground state is positive [41] and hence we use the pRNN wave function ansatz. The  $J_1$ - $J_2$  model with positive couplings, on the other hand, has a ground state endowed with a sign structure in the computational  $z$  basis, and thus we use a cRNN wave function ansatz.

In the following sections, we use 1D RNN wave functions to approximate the ground-state problem of the 1D TFIM and the 1D  $J_1$ - $J_2$  model, whereas we use both 1D and 2D pRNN wave functions in the case of the 2D TFIM.

#### A. 1D transverse field Ising model

To demonstrate the power of our proposed method, we use it to target the ground state of a TFIM in one dimension with  $N = 1000$  spins at the critical point  $h = 1$  using a pRNN wave function that has a single-layer RNN with 50 memory units. In Fig. 3(a), we show the evolution of the relative error,

$$\epsilon \equiv \frac{|E_{\text{RNN}} - E_{\text{DMRG}}|}{|E_{\text{DMRG}}|}, \quad (16)$$

and the energy variance per spin,

$$\sigma^2 \equiv \frac{\langle \hat{H}^2 \rangle - \langle \hat{H} \rangle^2}{N}, \quad (17)$$

as a function of the training step.  $E_{\text{DMRG}}$  is the ground-state energy as obtained from a density-matrix renormalization-group (DMRG) calculation [43,44], and can be considered exact in one dimension. We obtain very accurate results with a modest number of parameters ( $\approx 8000$ , see Appendix A). For comparison, the number of parameters of a restricted Boltzmann machine (RBM) [9] with one layer scales as  $MN$

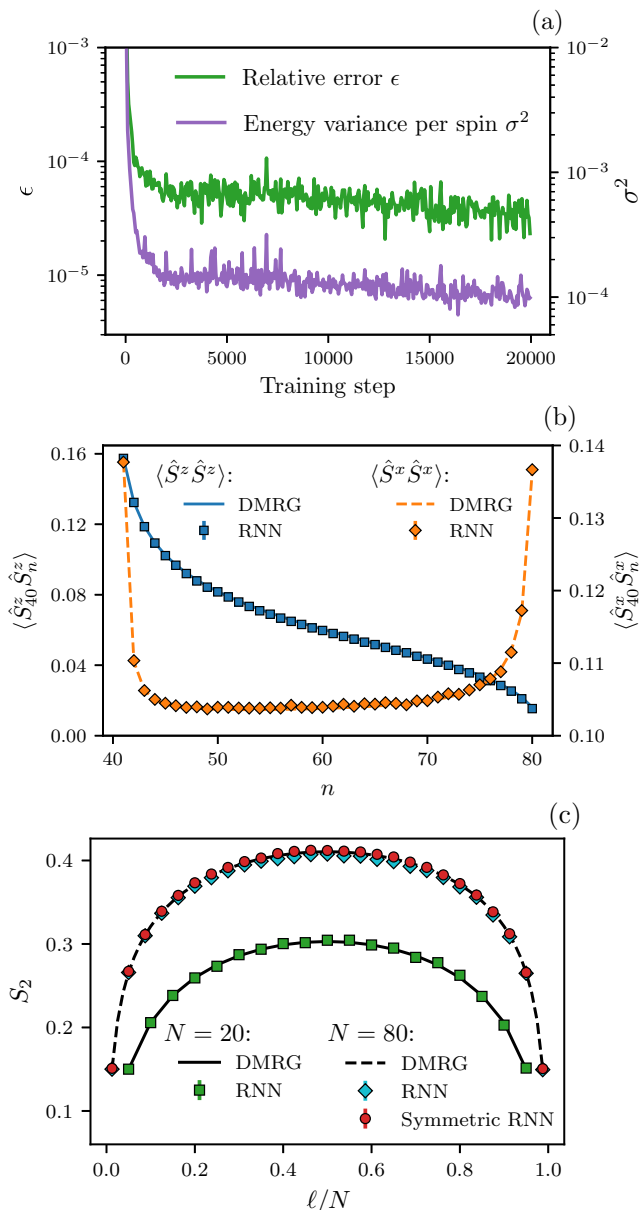


FIG. 3. Results for the pRNN wave function compared with DMRG when targeting the ground state of a 1D TFIM at the critical point. Our pRNN wave function has one layer with 50 units. (a) The relative error  $\epsilon$  and the energy variance per spin  $\sigma^2$  against the number of training steps (i.e., gradient descent steps) for  $N = 1000$  spins. We use only 200 samples per gradient step, which are enough to achieve convergence. (b) The two-point correlation function  $\langle \hat{S}_{40}^z \hat{S}_n^z \rangle$  along the  $x$  axis and  $z$  axis of the optimized pRNN wave function for sites  $n > 40$  using  $10^6$  samples. DMRG results are also shown for comparison. (c) The Rényi entropy  $S_2$  against the relative size of subregion  $A$  for system sizes  $N = 20$  and  $80$ . In both (b) and (c), the error bars are smaller than the data points.

with  $M$  the number of hidden units and  $N$  the number of physical spins. This scaling implies that the pRNN wave function here has the same number of variational parameters as an RBM with only eight hidden units.

While energies and variances give a quantitative indication of the quality of a variational wave function, correlation

functions provide a more comprehensive characterization. Indeed, correlation functions are at the heart of condensed-matter theory since many experimental probes in condensed-matter physics directly relate to measurements of correlation functions. Examples include inelastic scattering, which probes density-density correlation functions, and the Green's function, out of which important thermodynamic properties of a quantum system can be computed [45]. In Fig. 3(b) we compare the RNN results for the two-point correlation functions  $\langle \hat{S}_n^x \hat{S}_m^x \rangle$  and  $\langle \hat{S}_n^z \hat{S}_m^z \rangle$  with DMRG. Here, we see consistency between the RNN and the DMRG results.

Extracting entanglement entropy from many-body quantum systems is a central theme in condensed-matter physics, with entanglement entropy providing an additional window into the structure of complex quantum states of matter beyond what is seen from correlation functions. Of particular interest is the family of Rényi entropies of order  $\alpha$  of a reduced density matrix  $\rho$ :

$$S_\alpha(\rho) = \frac{1}{1-\alpha} \log(\text{Tr}\rho^\alpha). \quad (18)$$

$S_\alpha(\rho)$  encodes important nonlocal properties of quantum many-body systems such as topological entanglement, and contains information about universal properties of quantum phases such as the central charge  $c$  [46,47]. Due to their nonlocal character, extracting Rényi entropies from many-body quantum systems is notoriously difficult. Here, we use the so-called *replica trick* [48] to calculate the  $\alpha = 2$  Rényi entropy  $S_2(\rho)$  for RNN wave functions. The details of the implementation can be found in Appendix E. In Fig. 3(c), we show results for the Rényi entropy  $S_2(\rho_\ell)$  for two different system sizes  $N = 20, 80$  of the 1D TFIM.  $\rho_\ell$  here is the reduced density matrix on the first  $\ell$  sites of the spin chain, obtained by tracing out all sites  $n \in [\ell + 1, L]$  such that

$$\rho_\ell = \text{Tr}_{n \in [\ell+1, L]}(|\Psi\rangle\langle\Psi|). \quad (19)$$

Indeed, for both system sizes, Fig. 3(c) shows excellent agreement between the pRNN wave function estimation and the DMRG result. To improve the overall quality of the quantum state, we have enforced the parity symmetry on our pRNN wave function (see Appendix D 1), denoted by “Symmetric RNN” in Fig. 3(c). We observe that the symmetric pRNN wave function leads to a more accurate estimate of  $S_2(\rho_\ell)$  for  $N = 80$  sites.

### B. 1D $J_1$ - $J_2$ model

Moving beyond Stoquastic Hamiltonians, we now investigate the performance of RNN wave functions for a Hamiltonian the ground state of which has a sign structure in the computational basis, specifically the  $J_1$ - $J_2$  model in one dimension.

We use a variationally optimized deep cRNN wave function with three GRU layers, each with 100 memory units, to approximate the ground state of the  $J_1$ - $J_2$  model. The phase diagram of this model has been studied with DMRG [49], where it was found that the model exhibits a quantum phase transition at  $J_2^c = 0.241167 \pm 0.000005$  [50,51] from a critical Luttinger liquid phase for  $J_2 \leq J_2^c$  to a spontaneously dimerized gapped valence bond state phase for  $J_2 \geq J_2^c$ .

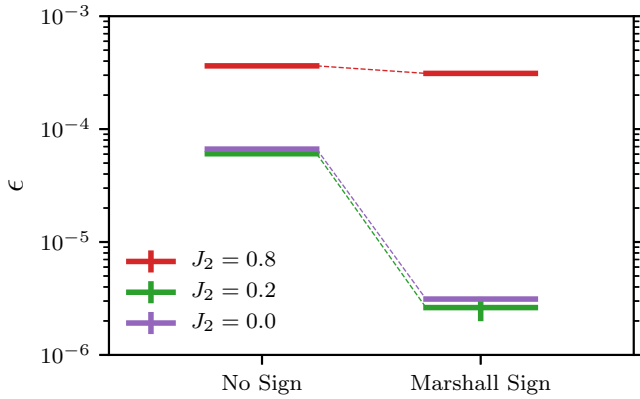


FIG. 4. The relative error (compared to DMRG) of the cRNN wave function trained on the 1D  $J_1$ - $J_2$  model with  $N = 100$  spins for different values  $J_2$ , both without a prior sign (represented by “No Sign”) and with a prior Marshall sign as in Eq. (20) (represented by “Marshall Sign”). We observe that applying a Marshall sign improves the accuracy.

We impose  $U(1)$  spin symmetry in the cRNN wave function (see Appendix D2), and target the ground state at four different points  $J_2 = 0.0, 0.2, 0.5, 0.8$ . Note that at  $J_2 = 0$  the Hamiltonian Eq. (15) can be made Stoquastic by a local unitary transformation that rotates every other spin by  $\pi$  around the  $z$  axis. The ground state can in this case be decomposed as [52]

$$\psi(\sigma) = (-1)^{M_A(\sigma)} \tilde{\psi}(\sigma), \quad (20)$$

where  $M_A(\sigma)$  is given by  $M_A(\sigma) = \sum_{i \in A} \sigma_i$  with  $\sigma_i \in \{0, 1\}$  [52] and  $\tilde{\psi}(\sigma)$  is the *positive* amplitude of the wave function. The set  $A$  comprises the sites belonging to the sublattice of all even (or all odd) sites in the lattice. The prefactor  $(-1)^{M_A(\sigma)}$  is known as the Marshall sign of the wave function [52]. For  $J_2 \neq 0$ , this decomposition is no longer exact, and  $\tilde{\psi}(\sigma)$  acquires a nontrivial sign structure. For finite  $J_2$  the decomposition in Eq. (20) can still be applied with the hope that the sign structure of  $\psi(\sigma)$  remains close to the Marshall sign [53].

In Fig. 4, we compare ground-state energies of the cRNN wave function trained on the 1D  $J_1$ - $J_2$  model with  $N = 100$  spins with and without applying a Marshall sign. For small values of  $J_2$ , we find a considerable improvement of the energies when applying the Marshall sign on top of the cRNN wave function. This observation highlights the importance of considering a prior “sign ansatz” to achieve better results. In the absence of a prior sign, the cRNN wave function can still achieve accurate estimations of the ground-state energies, showing that cRNN wave functions can recover some of the unknown sign structure of the ground state. For  $J_2 = 0.8$ , however, the improvement is less pronounced, which is expected due to the emergence of a second sign structure in the limit  $J_2 \rightarrow \infty$  (when the system decouples into two independent unfrustrated Heisenberg chains) [54,55], that is widely different from the Marshall sign in Eq. (20). We omit from Fig. 4 our results at the point  $J_2 = 0.5$ . In this case, the 1D  $J_1$ - $J_2$  model reduces to the Majumdar-Ghosh model, where the ground state is a product state of spin singlets, and

we find agreement with the exact ground-state energy within error bars when we apply an initial Marshall sign structure. We provide a summary of the cRNN wave function’s obtained values in Appendix F.

### C. 2D transverse field Ising model

Understanding strongly correlated quantum many-body systems in  $D > 1$  spatial dimensions is one of the central problems in condensed-matter physics. During the last decade, numerical approaches such as tensor networks [57–59], quantum Monte Carlo [13,60], and neural networks [9] have moved to the forefront of research in this area. Despite tremendous progress, however, solving correlated quantum many-body systems even in two dimensions remains a challenging problem. We now turn our attention to the application of our RNN wave function approach to the 2D quantum Ising model shown in Eq. (14) on a square lattice, a paradigmatic example of a strongly correlated quantum many-body system. This model has a quantum phase transition at a critical magnetic field  $h^c \approx 3.044$  that separates a magnetically ordered phase from a random paramagnet [61].

The simplest strategy for extending our approach to 2D geometries is to simply treat them as folded 1D chains, similar to the “snaking” approach used in 2D DMRG calculations [see Fig. 5(a)]. While this approach works quite well, it has the fundamental drawback that neighboring sites on the lattice can become separated in the 1D geometry. As a consequence, local correlations in the 2D lattice are mapped into nonlocal correlations in the 1D geometry, which can increase the complexity of the problem considerably. For example, 2D DMRG calculations are typically restricted to 2D lattices with small width  $L_y$ . This problem has led to the development of more powerful tensor network algorithms for 2D quantum systems such as projected entangled pair states (PEPS) [57].

An advantage of RNN wave functions is their flexibility in how hidden vectors are passed between units. To obtain an RNN wave function more suited to a 2D geometry, we modify the simple 1D approach outlined above by allowing hidden vectors to also be passed vertically, instead of only horizontally, as described in Appendix B. This modification is illustrated by the red arrows in Fig. 5(b). We refer to this geometry in the following discussions as a 2D RNN. We optimize the 2D pRNN wave function with a single-layer 2D vanilla RNN cell that has 100 memory units (i.e., with  $\approx 21\,000$  variational parameters) to approximate the ground state of the 2D quantum Ising model at  $h = 2, 3, 4$ . The training complexity of the 2D pRNN wave function is only quadratic in the number of memory units  $d_h$  (see Appendix B), which is very inexpensive compared to, e.g., the expensive variational optimization of PEPS, which scales as  $\chi^2 \bar{D}^6$  (where  $\bar{D}$  is the PEPS bond dimension and  $\chi$  is the bond dimension of the intermediate MPS) [62].

For comparison, we also optimize a deep 1D pRNN wave function architecture with three layers of stacked GRU cells, each with 100 memory units (i.e., with  $\approx 152\,000$  variational parameters) for the same values of the magnetic field  $h$ . In Fig. 5(c) we compare the obtained ground-state energies with results from 2D DMRG calculations (run on the same 1D geometry as for the 1D pRNN wave function) and the

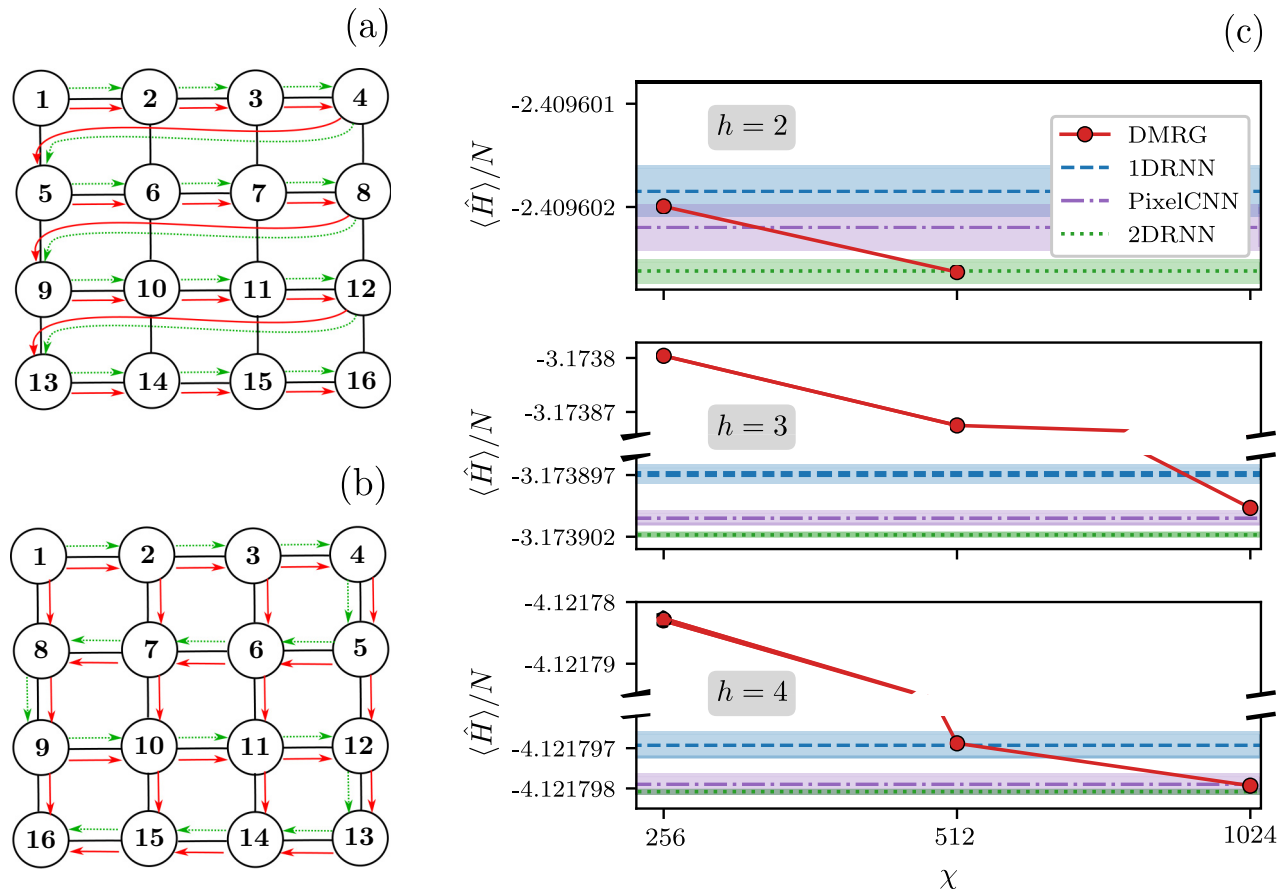


FIG. 5. (a) Autoregressive sampling path of 2D spin configurations using 1D RNN wave functions. The 2D configurations are generated through raster scanning, such that in order to generate spin  $\sigma_i$  one has to condition on the spins that are previously generated. (b) Autoregressive sampling path of 2D spin configurations using 2D RNN wave functions through a zigzag path, where each site receives two hidden states and two spins from the horizontal and the vertical neighbors that were previously generated. For both panels (a) and (b), the digits and the green dashed arrows indicate the sampling path, while the red arrows indicate how the hidden states are passed from one site to another. (c) A comparison of the variational energy per spin between a 2D pRNN wave function (labeled as 2DRNN), 1D pRNN wave function (labeled as 1DRNN), PixelCNN wave function [56], and DMRG with bond dimension  $\chi$  for the 2D TFIM on a system with  $L_x \times L_y = 12 \times 12$  spins. The shaded regions represent the error bars of each method. Note the broken y axis on the plots for  $h = 3$  and 4, denoting a change in scale between the upper and lower portions of the plots. These results show that 2D pRNN wave functions can achieve a performance comparable to PixelCNN wave functions and DMRG with a large bond dimension, while requiring only a fraction of their variational parameters.

PixelCNN architecture [63] (with  $\approx 800\,000$  variational parameters and results are taken from Ref. [56]). For the magnetic fields shown above and for large bond dimensions, we obtain excellent agreement between all four methods. This agreement is particularly remarkable given that the 2D pRNN wave function uses only about 0.03% of the variational parameters of the DMRG calculation with bond dimension  $\chi = 512$ , about 2.6% of the variational parameters of the PixelCNN wave function used in Ref. [56], and about 14% of the parameters used in the 1D pRNN architecture. A summary of our results in tabular form can be found in Appendix F.

#### D. Scaling of resources

The optimization results of our RNN wave function approach depend on several hyperparameters, including the number of memory units, the number of recurrent layers in deep architectures, and the number of samples used to obtain the gradient during an optimization step (see

Appendix C). Here, we investigate how the optimized energy and the energy variance per spin  $\sigma^2$  [see Eq. (17)] depend on these parameters. This energy variance per spin is an indicator of the quality of the optimized wave function, with exact eigenstates corresponding to  $\sigma^2 = 0$ . When targeting eigenstates, deviations from this value can be used to assess the quality of a variational wave function [13,64,65], as previously done in the case of matrix product state based techniques [66,67]. For variational approaches such as DMRG, one typically expects a nonzero value of  $\sigma^2$  that decreases when one increases the number of parameters (i.e., the expressivity) of the variational wave function. Since the number of variational parameters is directly related to the number of memory units of the pRNN wave function (see Appendix A), we study here the scaling of  $\sigma^2$  with the number of memory units.

In Fig. 6, we present the dependence of  $\sigma^2$  on the number of memory units for the 1D and 2D critical TFIMs. Fig. 6(a) shows results for  $\sigma^2$  for a 1D critical TFIM on three system

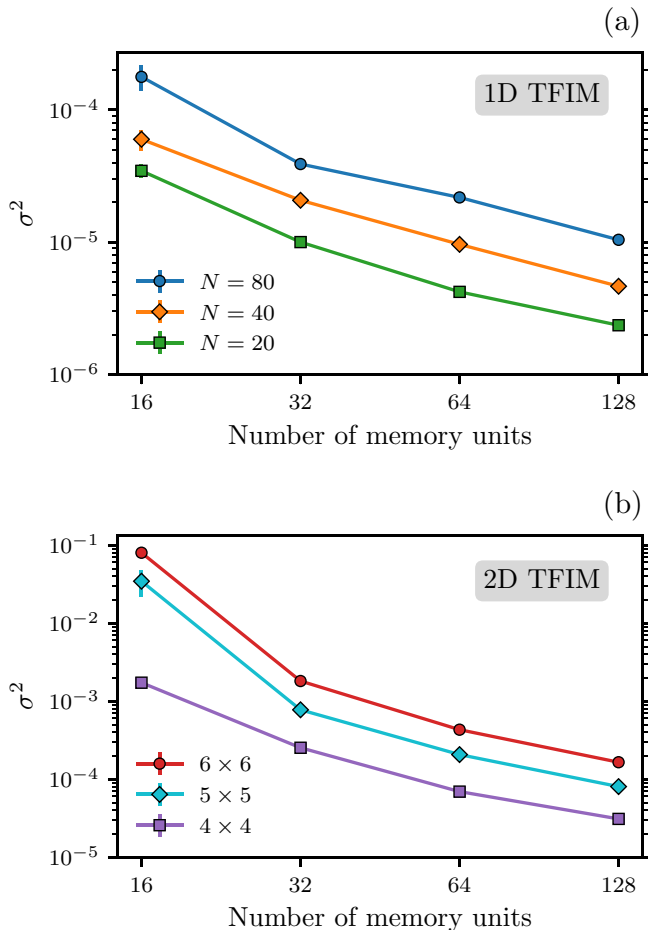


FIG. 6. The energy variance per spin against the number of memory units of a 1D pRNN wave function trained at the critical point of (a) the 1D TFIM and (b) the 2D TFIM. Both scalings show that we can systematically reduce the bias in the estimation of the ground-state energy.

sizes  $N = 20, 40$ , and  $80$ , and Fig. 6(b) shows results for the 2D TFIM on  $4 \times 4$ ,  $5 \times 5$  and  $6 \times 6$  square lattices. In all cases, we used a single-layer 1D pRNN wave function and 500 samples during optimization to compute estimates of the gradients. For each  $N$  we observe a systematic decrease of  $\sigma^2$  (i.e., an increase in quality of the wave function) as we increase the number of memory units.

In Appendix G, we study the dependence of  $\sigma^2$  on both the number of samples and the number of layers in the pRNN wave function for a critical 1D TFIM. We observe only a weak dependence on both parameters. The weak dependence on the number of samples suggests that optimizing the RNN wave functions with noisy gradients does not significantly impact the results of the optimization procedure, and yields accurate estimations of the ground state and its energy. From the weak dependence on the number of layers we conclude that shallow RNNs with a sufficient number of memory units have enough expressivity and that deep architectures do not seem to be beneficial from an accuracy point of view. However, deeper networks could have potential ramifications regarding memory usage and training speed when it comes to training a large number of variational parameters, as shallow RNNs with

a large number of memory units are equivalent in terms of number of parameters to deep RNNs with a smaller number of memory units. We also note that adding residual connections between layers [68] and dilated connections between RNN cells [69] to deep RNNs, which we leave for future investigations, might change our previous conclusions and make deep RNNs more beneficial compared to shallow RNNs.

#### IV. CONCLUSIONS AND OUTLOOK

We have introduced recurrent neural network wave functions, a variational ansatz for quantum many-body systems, which we use to approximate ground-state energies, correlation functions, and entanglement of many-body Hamiltonians of interest to condensed-matter physics. We find that RNN wave functions are competitive with state-of-the-art methods such as DMRG and PixelCNN wave functions [56], performing particularly well on the task of finding the ground state of the transverse field Ising model in two dimensions. By increasing the number of memory units in the RNN, the error in our results can be systematically reduced. We have shown furthermore that we can accurately model ground states endowed with a sign structure using a complex recurrent neural network (cRNN) wave function ansatz. Here, accuracy can be improved by introducing an ansatz sign structure and by enforcing symmetries such as  $U(1)$  symmetry. The autoregressive nature of RNN wave functions makes it possible to directly generate independent samples, in contrast to methods based on Markov chain sampling, which are often plagued by long autocorrelation times that affect the optimization and the accurate estimation of correlation functions in a variational ansatz. Thanks to weight sharing among lattice sites, RNN wave functions provide very compact yet expressive representations of quantum states, while retaining the ability to easily train with millions of variational parameters, as opposed to, e.g., restricted Boltzmann machines [9]. We expect that future work incorporating additional numerical techniques such as attention [25,70] and higher-order optimization [13,71] will make RNN wave functions a highly competitive tool for simulating quantum many-body systems, with applications to material science, quantum chemistry [72], quantum computation [73], and beyond.

Our code is made publicly available in Ref. [74]. The hyperparameters we use are given in Appendix H.

*Note added.* Recently, a paper on recurrent neural network wave functions [75] appeared after the publication of this manuscript.

#### ACKNOWLEDGMENTS

We acknowledge Di Luo for his generous comments, which were extremely helpful. We thank Estelle Inack, Dan Sehayek, Amine Natick, Matthew Beach, Bohdan Kulchytskyy, Florian Hopfmueller, Roeland Wiersema, Giuseppe Carleo, and Noam Wies for useful discussions and insights. We also thank our anonymous reviewer and our editor for their efforts and their insightful suggestions. M.H. acknowledges support by the Ecole Normale Supérieure de Lyon. M.G. acknowledges support by the Simons Foundation (Many Electron Collaboration). R.G.M. acknowledges



support from the Natural Sciences and Engineering Research Council of Canada (NSERC), a Canada Research Chair, the Shared Hierarchical Academic Research Computing Network (SHARCNET), and Compute Canada. J.C. acknowledges support from NSERC, SHARCNET, Compute Canada, and the Canadian institute for advanced research (CIFAR) AI chair program. Computer simulations were also made possible thanks to the Vector Institute computing cluster and Google Colaboratory. This research was supported in part by the National Science Foundation under Grant No. NSF PHY-1748958. It was also supported in part by the Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported in part by the Government of Canada through Innovation, Science and Economic Development Canada (ISED) and by the Province of Ontario through the Ministry of Economic Development, Job Creation, and Trade.

### APPENDIX A: GATED RECURRENT NEURAL NETWORKS

We use the GRU model introduced in Ref. [40], which processes the spin configurations  $\sigma$  as

$$\begin{aligned} \mathbf{u}_n &= \text{sig}(W_u[\mathbf{h}_{n-1}; \sigma_{n-1}] + \mathbf{b}_u), \\ \mathbf{r}_n &= \text{sig}(W_r[\mathbf{h}_{n-1}; \sigma_{n-1}] + \mathbf{b}_r), \\ \tilde{\mathbf{h}}_n &= \tanh(W_c[\mathbf{r}_n \odot \mathbf{h}_{n-1}; \sigma_{n-1}] + \mathbf{b}_c), \\ \mathbf{h}_n &= (1 - \mathbf{u}_n) \odot \mathbf{h}_{n-1} + \mathbf{u}_n \odot \tilde{\mathbf{h}}_n, \end{aligned} \quad (\text{A1})$$

where sig and tanh represent the sigmoid and hyperbolic tangent activation functions, respectively. Thus, the hidden vector  $\mathbf{h}_n$  is updated through an interpolation between the previous hidden state  $\mathbf{h}_{n-1}$  and a candidate hidden state  $\tilde{\mathbf{h}}_n$ . The update gate  $\mathbf{u}_n$  decides to what extent the contents of the hidden state are modified, and depends on how relevant the input  $\sigma_{n-1}$  is to the prediction (Softmax layer). The symbol  $\odot$  denotes the pointwise (Hadamard) product. The reset gate modeled by the vector  $\mathbf{r}_n$  is such that if the  $i$ th component  $r_n$  is close to zero it cancels out the  $i$ th component of the hidden vector state  $\mathbf{h}_{n-1}$ , effectively making the GRU “forget” part of the sequence that has already been encoded in the state vector  $\mathbf{h}_{n-1}$ .

The weights matrices  $W_{u,r,c}$  and the bias vectors  $\mathbf{b}_{u,r,c}$  parametrize the GRU and are optimized using energy minimization as described in Appendix C. The GRU transformations in Eq. (A1) are depicted graphically in Fig. 7.

To take advantage of the GPU speed up, we use instead the cuDNN variant of GRUs implemented in TENSORFLOW [76], with

$$\begin{aligned} \mathbf{u}_n &= \text{sig}(W_u[\mathbf{h}_{n-1}; \sigma_{n-1}] + \mathbf{b}_u), \\ \mathbf{r}_n &= \text{sig}(W_r[\mathbf{h}_{n-1}; \sigma_{n-1}] + \mathbf{b}_r), \\ \mathbf{h}'_n &= W_c^{(1)}\mathbf{h}_{n-1} + \mathbf{b}_c^{(1)}, \\ \tilde{\mathbf{h}}_n &= \tanh(W_c^{(2)}\sigma_{n-1} + \mathbf{r}_n \odot \mathbf{h}'_n + \mathbf{b}_c^{(2)}), \\ \mathbf{h}_n &= (1 - \mathbf{u}_n) \odot \mathbf{h}_{n-1} + \mathbf{u}_n \odot \tilde{\mathbf{h}}_n, \end{aligned} \quad (\text{A2})$$

which differs slightly from the above implementation of traditional GRU cells [77].

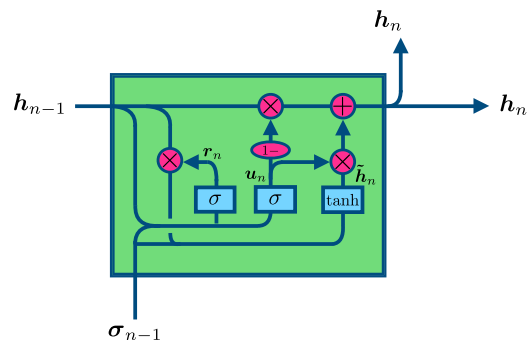


FIG. 7. Graphical representation of the gated recurrent unit cell described in Eq. (A1). The magenta circles/ellipses represent pointwise operations such as vector addition or multiplication. The blue rectangles represent neural network layers labeled by the nonlinearity we use. Merging lines denote vector concatenation and forking lines denote a copy operation. The sigmoid activation function is represented by  $\sigma$ .

Provided that the dimensions of the hidden state  $\mathbf{h}_{n-1}$  and input  $\sigma_{n-1}$  are  $d_h$  and  $d_v$ , respectively, then the dimensions of the variational parameters of a GRU as in Eq. (A2) are (1)  $\dim(W_{u,r}) = d_h \times (d_h + d_v)$ , (2)  $\dim(\mathbf{b}_{u,r}) = d_h$ , (3)  $\dim(W_c^{(1)}) = d_h \times d_h$ , (4)  $\dim(W_c^{(2)}) = d_h \times d_v$ , and (5)  $\dim(\mathbf{b}_c^{(1,2)}) = d_h$ . The new hidden state  $\mathbf{h}_n$  is fed into a Softmax layer to infer conditional probabilities, such that

$$\mathbf{y}_n^{(1)} = \text{Softmax}(U^{(1)}\mathbf{h}_n + \mathbf{c}^{(1)}),$$

and also into a Softsign layer to infer the phases as

$$\mathbf{y}_n^{(2)} = \pi \text{Softsign}(U^{(2)}\mathbf{h}_n + \mathbf{c}^{(2)}).$$

We require the outputs  $\mathbf{y}_n^{(1,2)}$  to have dimension  $d_v$ , so that each element of  $\mathbf{y}_n^{(1)}$  represents the conditional probability of sampling a value for the next spin  $\sigma_n \in \{0, 1, \dots, d_v - 1\}$ , and that each element of  $\mathbf{y}_n^{(2)}$  corresponds to the phase of the chosen spin  $\sigma_n$ . Thus, the dimensions of the parameters introduced in the Softmax/Softsign layer are (1)  $\dim(U^{(1,2)}) = d_v \times d_h$  and (2)  $\dim(\mathbf{c}^{(1,2)}) = d_v$ . The same reasoning can be also applied to determine the dimensions of the variational parameters of 2D vanilla RNNs presented in Appendix B.

### APPENDIX B: TWO-DIMENSIONAL RECURRENT NEURAL NETWORK WAVE FUNCTIONS

Standard RNN architectures are inherently one dimensional. However, most interesting quantum many-body systems live in higher dimensions. By taking inspiration from Refs. [21,78], we generalize one-dimensional RNNs to multi-dimensional RNN wave functions. In particular, we generalize to 2D vanilla RNNs that are more suitable to simulating two-dimensional square lattices than one-dimensional RNNs, which map two-dimensional lattice configurations to one-dimensional configurations and do not necessarily encode spatial information about neighboring sites in a plausible manner.

The main idea behind the implementation of 2D RNNs [21] is to replace the single hidden state that is passed from one site to another by two hidden states, with each one corresponding to the state of a neighboring site (vertical and horizontal) and hence respecting the 2D geometry of the problem. To do so,

we change the one-dimensional recursion relation in Eq. (2) to the two-dimensional recursion relation

$$\mathbf{h}_{i,j} = f(W^{(h)}[\mathbf{h}_{i-1,j}; \boldsymbol{\sigma}_{i-1,j}] + W^{(v)}[\mathbf{h}_{i,j-1}; \boldsymbol{\sigma}_{i,j-1}] + \mathbf{b}), \quad (\text{B1})$$

where  $\mathbf{h}_{i,j}$  is the hidden state at site  $(i, j)$ ,  $W^{(v,h)}$  are weight matrices, and  $\mathbf{b}$  is a bias. Here  $f$  is a nonlinear activation function chosen to be equal to the exponential linear unit (ELU) defined as

$$\text{ELU}(x) = \begin{cases} x, & \text{if } x \geq 0, \\ \exp(x) - 1, & \text{if } x < 0. \end{cases}$$

The cost of computing a new hidden state  $\mathbf{h}_{i,j}$  is quadratic in the size of the hidden state (number of memory units  $d_h$ ), and the cost of computing the gradients with respect to the variational parameters of the 2D RNN remains unchanged. This property allows us to train 2D RNNs with a relatively large  $d_h$ .

To initialize the 2D RNN, we choose  $\mathbf{h}_{i,0}$ ,  $\boldsymbol{\sigma}_{i,0}$  and  $\mathbf{h}_{0,j}$ ,  $\boldsymbol{\sigma}_{0,j}$  to be null vectors. Once  $\mathbf{h}_{i,j}$  is computed, we apply the same scheme as in Sec. II B to sample a spin  $\sigma_{i,j}$ . The scheme for computing positive or complex amplitudes from Sec. II B remains the same.

We note that generalization to higher dimensions, to other lattices, as well as to other types of RNN architectures can be done by taking inspiration from this scheme. For instance, using LSTMs [20], GRUs [40], or transformers [25] instead of vanilla RNNs in two dimensions is expected to make a significant improvement. We also expect that using multiplicative interactions [79] might increase the expressiveness of 2D RNNs as compared to the additive interactions in Eq. (B1).

### APPENDIX C: VARIATIONAL MONTE CARLO AND VARIANCE REDUCTION

The main goal of Variational Monte Carlo (VMC) is to iteratively optimize an ansatz wave function to approximate, e.g., ground states of local Hamiltonians. VMC starts from a suitable *trial wave function*  $|\Psi_\lambda\rangle$  that incorporates the variational degrees of freedom of the approach.  $|\Psi_\lambda\rangle$  could be, for example, an MPS wave function [80], in which case the free parameters are the MPS matrices. Crucially, the ansatz  $|\Psi_\lambda\rangle$  has to allow for *efficient sampling* from the square of the amplitudes of  $|\Psi_\lambda\rangle$ . In this paper, we choose RNN wave functions, described in Sec. II B, to parametrize the trial wave function  $|\Psi_\lambda\rangle$  for a VMC optimization of ground states.

The aim of the VMC optimization is to minimize the expectation value of the energy

$$E \equiv \frac{\langle \Psi_\lambda | \hat{H} | \Psi_\lambda \rangle}{\langle \Psi_\lambda | \Psi_\lambda \rangle} \quad (\text{C1})$$

when given a family of states  $|\Psi_\lambda\rangle$ . This minimization is carried out using the gradient descent method or any of its variants. Since the RNN wave function is normalized such that  $\langle \Psi_\lambda | \Psi_\lambda \rangle = 1$ , the expectation value in Eq. (C1) can be

written as

$$\begin{aligned} E &= \langle \Psi_\lambda | H | \Psi_\lambda \rangle = \sum_{\boldsymbol{\sigma}} |\psi_\lambda(\boldsymbol{\sigma})|^2 \sum_{\boldsymbol{\sigma}'} H_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} \frac{\psi_\lambda(\boldsymbol{\sigma}')}{\psi_\lambda(\boldsymbol{\sigma})} \\ &\equiv \sum_{\boldsymbol{\sigma}} |\psi_\lambda(\boldsymbol{\sigma})|^2 E_{\text{loc}}(\boldsymbol{\sigma}) \\ &\approx \frac{1}{N_S} \sum_{\boldsymbol{\sigma} \sim |\psi_\lambda(\boldsymbol{\sigma})|^2} E_{\text{loc}}(\boldsymbol{\sigma}), \end{aligned} \quad (\text{C2})$$

which represents a sample average of the local energy  $E_{\text{loc}}(\boldsymbol{\sigma})$ . Denoting  $\lambda_i$  to be the real variational parameters of  $|\Psi_\lambda\rangle$ , the gradients  $\partial_{\lambda_i} E$  can be similarly written as

$$\partial_{\lambda_i} E = \sum_{\boldsymbol{\sigma}} |\psi_\lambda(\boldsymbol{\sigma})|^2 \frac{\partial_{\lambda_i} \psi_\lambda^*(\boldsymbol{\sigma})}{\psi_\lambda^*(\boldsymbol{\sigma})} E_{\text{loc}}(\boldsymbol{\sigma}) + \text{c.c.} \quad (\text{C3})$$

An optimization step consists of drawing  $N_S$  samples  $\{\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}, \dots, \boldsymbol{\sigma}^{(N_S)}\}$  from  $|\psi_\lambda(\boldsymbol{\sigma})|^2$  autoregressively using the RNN wave function, and then computing  $\partial_{\lambda_i} E$  from Eq. (C3) as

$$\partial_{\lambda_i} E \approx \frac{2}{N_S} \text{Re} \left( \sum_{i=1}^{N_S} \frac{\partial_{\lambda_i} \psi_\lambda^*(\boldsymbol{\sigma}^{(i)})}{\psi_\lambda^*(\boldsymbol{\sigma}^{(i)})} E_{\text{loc}}(\boldsymbol{\sigma}^{(i)}) \right), \quad (\text{C4})$$

using automatic differentiation [81] and updating the parameters (if using gradient descent) according to

$$\lambda_i \leftarrow \lambda_i - \alpha \partial_{\lambda_i} E \quad (\text{C5})$$

with a small learning rate  $\alpha$ . Instead of this simple gradient descent rule, we use the ADAM optimizer [82] to implement the gradient updates. We found that the latter gives better results compared to the simple gradient descent optimization shown in Eq. (C5) and without having to carefully tune the learning rate  $\alpha$ .

We note that the stochastic evaluation of the gradients in Eq. (C4) tends to carry noise that increases their variances [83,84]. Such high variances tend to slow down the convergence to the ground-state energy. We propose to cure this limitation by introducing a new term in Eq. (C4) that helps reduce the variance of the gradients by approximating

$$\begin{aligned} \partial_{\lambda_i} E &\approx \frac{2}{N_S} \text{Re} \left\{ \sum_{i=1}^{N_S} \frac{\partial_{\lambda_i} \psi_\lambda^*(\boldsymbol{\sigma}^{(i)})}{\psi_\lambda^*(\boldsymbol{\sigma}^{(i)})} [E_{\text{loc}}(\boldsymbol{\sigma}^{(i)}) - E] \right\} \\ &= \frac{2}{N_S} \text{Re} \left\{ \sum_{i=1}^{N_S} \partial_{\lambda_i} \log \psi_\lambda^*(\boldsymbol{\sigma}^{(i)}) [E_{\text{loc}}(\boldsymbol{\sigma}^{(i)}) - E] \right\}, \end{aligned} \quad (\text{C6})$$

and we show below that this approximation does not introduce a bias. This new term is useful for reducing the uncertainty in the gradient estimation, as in the limit where  $E_{\text{loc}}(\boldsymbol{\sigma}^{(i)}) \approx E$  near convergence the variance of the gradients  $\partial_{\lambda_i} E$  goes to zero as opposed to the nonzero variance of the gradients in Eq. (C4). As a consequence, a stable convergence to the ground state is achieved as confirmed by our experiments. This idea is similar in spirit to control variate methods in Monte Carlo [83] and to baseline methods in reinforcement learning [85].

To show that the term we add in Eq. (C6) does not bias the true gradients in Eq. (C3), it suffices to prove that

$$\text{Re}\{\langle \partial_{\lambda_i} \log[\psi_{\lambda}^*(\sigma)] E \rangle\} = 0, \quad (\text{C7})$$

where  $\langle \dots \rangle$  denotes the statistical average over the probability distribution  $|\psi_{\lambda}|^2$ . To prove this expression, we write  $\psi_{\lambda}(\sigma) = \sqrt{P_{\lambda}(\sigma)} \exp[i\phi_{\lambda}(\sigma)]$ , which implies that

$$\log[\psi_{\lambda}^*(\sigma)] = \frac{1}{2} \log[P_{\lambda}(\sigma)] - i\phi_{\lambda}(\sigma),$$

and hence

$$\begin{aligned} & \text{Re}\{\langle \partial_{\lambda_i} \log[\psi_{\lambda}^*(\sigma)] E \rangle\} \\ &= \frac{1}{2} \langle \partial_{\lambda_i} \log[P_{\lambda}(\sigma)] \rangle \text{Re}(E) + \langle \partial_{\lambda_i} \phi_{\lambda}(\sigma) \rangle \text{Im}(E). \end{aligned} \quad (\text{C8})$$

To show that  $\langle \partial_{\lambda_i} \log[P_{\lambda}(\sigma)] \rangle = 0$  [85,86], we write

$$\begin{aligned} \langle \partial_{\lambda_i} \log[P_{\lambda}(\sigma)] \rangle &= \sum_{\sigma} P_{\lambda}(\sigma) \partial_{\lambda_i} \log[P_{\lambda}(\sigma)] \\ &= \sum_{\sigma} P_{\lambda}(\sigma) \frac{\partial_{\lambda_i} P_{\lambda}(\sigma)}{P_{\lambda}(\sigma)} \\ &= \partial_{\lambda_i} \sum_{\sigma} P_{\lambda}(\sigma) \\ &= \partial_{\lambda_i} 1 = 0, \end{aligned}$$

where the fact that the RNN wave function is normalized justifies the transition from the third line to the fourth line.

From here, it suffices to show that  $\langle \partial_{\lambda_i} \phi_{\lambda}(\sigma) \rangle \text{Im}(E) = 0$ . Since the Hamiltonian  $\hat{H}$  is Hermitian, the expectation value  $E$  is real and hence  $\text{Im}(E) = 0$ . We therefore arrive at Eq. (C7).

## APPENDIX D: IMPLEMENTING SYMMETRIES

### 1. Imposing discrete symmetries

Inspired by Refs. [32,56], we propose to implement discrete symmetries in a similar fashion for RNN wave functions without spoiling their autoregressive nature.

Assuming that a Hamiltonian  $\hat{H}$  has a symmetry under discrete transformations  $\mathcal{T}$ , its ground state

$$|\Psi_G\rangle = \sum_{\sigma} \psi_G(\sigma) |\sigma\rangle$$

is an eigenvector of the symmetry transformation  $\mathcal{T}$ . The ground state transforms as  $\psi_G(\mathcal{T}\sigma) = \omega_{\mathcal{T}} \psi_G(\sigma)$  where  $\omega_{\mathcal{T}}$  is an eigenvalue with module 1, that is independent of the choice of  $\sigma$ . This expression implies that the transformation  $\mathcal{T}$  changes the ground state with only a global phase term that does not affect the probability distribution, and changes the sign structure with a global phase term. It is thus desirable that the RNN wave function also has this symmetry.

To enforce a discrete symmetry  $\{\mathcal{T}\}$  on an RNN wave function  $|\Psi_{\lambda}\rangle$ , we propose the following scheme.

(1) Generate a sample  $\sigma$  autoregressively from the RNN wave function.

(2) Sample with a probability  $1/\text{Card}(G)$  a transformation  $\mathcal{T}$  from the symmetry transformation group  $G = \{\mathbb{1}, \mathcal{T}_1, \dots\}$  that leaves the Hamiltonian  $\hat{H}$  invariant, and apply the transformation  $\mathcal{T}$  to  $\sigma$ .

(3) Assign to the spin configuration  $\tilde{\sigma} = \mathcal{T}\sigma$  the amplitude  $\psi_{\lambda}(\tilde{\sigma}) = \sqrt{P_{\lambda}(\tilde{\sigma})} \exp[i\phi_{\lambda}(\tilde{\sigma})]$ , such that

$$\begin{aligned} P_{\lambda}(\tilde{\sigma}) &= \frac{1}{\text{Card}(G)} \left[ \sum_{\tilde{\mathcal{T}} \in G} P_{\lambda}(\tilde{\mathcal{T}}\sigma) \right], \\ \phi_{\lambda}(\tilde{\sigma}) &= \text{Arg} \left\{ \omega_{\mathcal{T}} \sum_{\tilde{\mathcal{T}} \in G} \exp[i\phi_{\lambda}(\tilde{\mathcal{T}}\sigma)] \right\}, \end{aligned}$$

where  $P_{\lambda}(\tilde{\mathcal{T}}\sigma)$  is a probability generated by the Softmax layer and  $\phi_{\lambda}(\tilde{\mathcal{T}}\sigma)$  is a phase generated by the Softsign layer, as explained in Sec. II B.

If the ground state is positive [41], we use the same algorithm but only symmetrize the probability  $P_{\lambda}$ , without having to worry about symmetrizing the phase  $\phi_{\lambda}$ .

For concreteness, we illustrate the algorithm above with ‘‘symmetric RNNs’’ that have a built-in parity symmetry. We use this architecture in Sec. III A to get a more accurate estimate of the ground state of the 1D TFIM that also obeys a parity symmetry. Indeed, symmetric RNNs show an improvement over ordinary pRNN wave functions on the task of estimating the second Rényi entropy (see Appendix E). Symmetric RNNs can be implemented using the following procedure.

- (1) Sample each configuration  $\sigma$ .
- (2) Choose to apply or to not apply the parity transformation  $\hat{P}$  on  $\sigma$  with a probability  $1/2$ .
- (3) Assign to  $\sigma$  the probability:

$$P = \frac{[P_{\lambda}(\sigma) + P_{\lambda}(\hat{P}\sigma)]}{2}.$$

We also emphasize the possibility of carefully designing RNN wave functions to impose discrete symmetries, without using the symmetrization scheme above and which we leave for future investigations.

### 2. Imposing zero magnetization

Since the ground state of the  $J_1$ - $J_2$  model has zero magnetization, i.e., a  $U(1)$  symmetry [52,87], it is helpful to enforce this constraint on our RNN wave functions to get accurate estimations of the ground-state energy. To do so, we propose an efficient way to generate samples with zero magnetization while maintaining the autoregressive property of the RNN wave function. The procedure effectively applies a projector  $\mathcal{P}_{S_z=0}$  to the original state, which restricts the RNN wave function to the subspace of configurations with zero magnetization. This procedure avoids generating a large number of samples and discarding the ones that have nonzero magnetization.

The condition of zero magnetization implies that the number of up spins should be equal to the number of down spins. To satisfy this constraint, we utilize the following algorithm.

(1) Sample autoregressively the first half of the spin configuration  $(\sigma_1, \sigma_2, \dots, \sigma_{N/2})$ .

(2) At each step  $i > N/2$ , (i) generate the output of the RNN wave function,  $y_i = (\psi_i^{\text{down}}, \psi_i^{\text{up}})$ , where  $\psi_i^{\text{down}}$  and  $\psi_i^{\text{up}}$  are both nonzero and their modules squared sum to 1; and (ii)

define the following amplitudes:

$$a_i = \psi_i^{\text{down}} \times \Xi\left(\frac{N}{2} - N_{\text{down}}(i)\right),$$

$$b_i = \psi_i^{\text{up}} \times \Xi\left(\frac{N}{2} - N_{\text{up}}(i)\right),$$

where

$$\Xi(x) \equiv \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0, \end{cases}$$

and

$$N_{\text{down}}(i) = \text{Card}(\{j / \sigma_j = 0 \text{ and } j < i\}),$$

$$N_{\text{up}}(i) = \text{Card}(\{j / \sigma_j = 1 \text{ and } j < i\}).$$

In words,  $N_{\text{up}}(i)/N_{\text{down}}(i)$  is the number of up/down spins generated before step  $i$ .

(3) Sample  $\sigma_i$  from  $|\tilde{y}_i|^2$ , where

$$\tilde{y}_i = \frac{1}{\sqrt{a_i^2 + b_i^2}}(a_i, b_i),$$

which is normalized, i.e.,  $\|\tilde{y}_i\|_2 = 1$ .

Using this algorithm, it is clear that the RNN wave function generates a spin configuration that has the same number of up spins and down spins, and hence a zero magnetization. In fact, at each step  $i > N/2$ , the function  $\Xi$  assigns a zero amplitude for the next spin  $\sigma_i$  to be spin up if  $N_{\text{up}}(i) = N/2$  or to be spin down if  $N_{\text{down}}(i) = N/2$ .

Interestingly enough, our scheme does not spoil the normalization of the RNN wave function as the new conditional probabilities  $|\tilde{y}_i|^2$  are also normalized. We also note that this algorithm preserves the autoregressive property of the original RNN wave function and can also be parallelized. Moreover, this scheme can be easily extended to the generation of samples with a nonzero fixed magnetization, which is useful when considering the problem of finding states that live in a nonzero fixed magnetization sector.

## APPENDIX E: RÉNYI ENTROPIES

Given a quantum system with a spatial bipartition  $(A, B)$ , one can write the RNN wave function  $|\Psi_\lambda\rangle$  as

$$|\Psi_\lambda\rangle = \sum_{\sigma_A, \sigma_B} \psi_\lambda(\sigma_A \sigma_B) |\sigma_A \sigma_B\rangle,$$

where  $\sigma_{A/B}$  denotes the spin configuration that lives in the partition  $A/B$  and  $\sigma_A \sigma_B$  stands for a concatenation of  $\sigma_A$  and  $\sigma_B$ .

The  $\alpha$ -Rényi entropy between region  $A$  and  $B$  is given by

$$S_\alpha(A) = \frac{1}{1-\alpha} \log(\text{Tr} \rho_A^\alpha), \quad (\text{E1})$$

where  $\rho_A = \text{Tr}_B |\Psi_\lambda\rangle \langle \Psi_\lambda|$  and  $\alpha$  is an integer [48]. To estimate these entropies, we use the so-called replica trick [48], where we consider the action of the  $\text{Swap}_A$  operator on the two copies of the RNN wave function, which swaps the spins in the region  $A$  between the two copies (as demonstrated

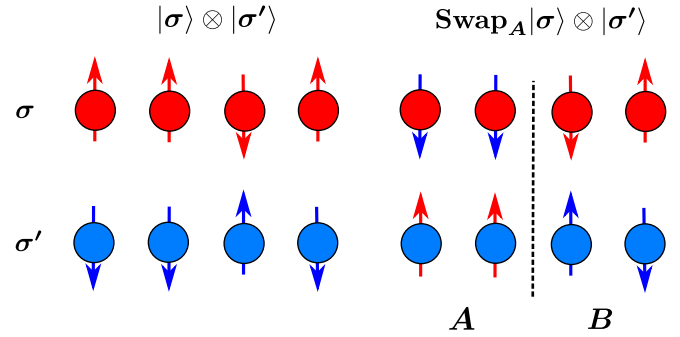


FIG. 8. The Swap operator acting on the tensor product of two samples  $\sigma$  and  $\sigma'$ .

in Fig. 8) such that

$$\text{Swap}_A |\Psi_\lambda\rangle \otimes |\Psi_\lambda\rangle = \sum_{\sigma, \tilde{\sigma}} \psi_\lambda(\sigma_A \sigma_B) \psi_\lambda(\tilde{\sigma}_A \tilde{\sigma}_B) |\tilde{\sigma}_A \sigma_B\rangle \otimes |\sigma_A \tilde{\sigma}_B\rangle. \quad (\text{E2})$$

The expectation value of  $\text{Swap}_A$  in the double copy of the RNN wave function “ $|\Psi_\lambda\rangle \otimes |\Psi_\lambda\rangle$ ” is given by [42,48]

$$\langle \text{Swap}_A \rangle = \sum_{\sigma, \tilde{\sigma}} \psi_\lambda^*(\sigma_A \sigma_B) \psi_\lambda^*(\tilde{\sigma}_A \tilde{\sigma}_B) \psi_\lambda(\tilde{\sigma}_A \sigma_B) \psi_\lambda(\sigma_A \tilde{\sigma}_B) = \text{Tr} \rho_A^2 = \exp[-S_2(A)]. \quad (\text{E3})$$

Hence, by calculating the expectation of the value of the Swap operator in the double copy of the RNN wave function, we can access the second Rényi entropy. Interestingly, the Rényi entropies  $S_\alpha$  have been shown to encode similar properties independently of  $\alpha$  [46,48].

Although an exact evaluation of Eq. (E3) is numerically intractable, we can use importance sampling to estimate it [48] as

$$\langle \text{Swap}_A \rangle = \sum_{\sigma, \tilde{\sigma}} |\psi_\lambda(\sigma_A \sigma_B)|^2 |\psi_\lambda(\tilde{\sigma}_A \tilde{\sigma}_B)|^2 \frac{\psi_\lambda(\tilde{\sigma}_A \sigma_B) \psi_\lambda(\sigma_A \tilde{\sigma}_B)}{\psi_\lambda(\sigma_A \sigma_B) \psi_\lambda(\tilde{\sigma}_A \tilde{\sigma}_B)},$$

$$\approx \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{\psi_\lambda(\tilde{\sigma}_A^{(i)} \sigma_B^{(i)}) \psi_\lambda(\sigma_A^{(i)} \tilde{\sigma}_B^{(i)})}{\psi_\lambda(\sigma_A^{(i)} \sigma_B^{(i)}) \psi_\lambda(\tilde{\sigma}_A^{(i)} \tilde{\sigma}_B^{(i)})}. \quad (\text{E4})$$

Using this trick, for the system sizes studied in this paper we only have to generate two sets of exact samples  $\{\sigma^{(i)}\}_{i=1}^{N_s}$  and  $\{\tilde{\sigma}^{(i)}\}_{i=1}^{N_s}$  independently from  $|\psi_\lambda|^2$  without having to use the improved ratio trick [48]. By defining

$$\text{Swap}_A^{(i)} \equiv \frac{\psi_\lambda(\tilde{\sigma}_A^{(i)} \sigma_B^{(i)}) \psi_\lambda(\sigma_A^{(i)} \tilde{\sigma}_B^{(i)})}{\psi_\lambda(\sigma_A^{(i)} \sigma_B^{(i)}) \psi_\lambda(\tilde{\sigma}_A^{(i)} \tilde{\sigma}_B^{(i)})},$$

the statistical error on the estimation of the Rényi-2 entropy can be calculated as

$$\epsilon = \frac{1}{\langle \text{Swap}_A \rangle} \sqrt{\frac{\text{var}(\{\text{Swap}_A^{(i)}\})}{N_s}}.$$

For the estimation of the Rényi-2 entropy for the 1D TFIM in this paper, we use  $N_s = 2 \times 10^6$  samples from a trained

TABLE I. Energy per spin values for the 1D  $J_1$ - $J_2$  model. We consider a cRNN wave function with two different methods of training (with no initial sign structure and with a background Marshall sign) and compare with results from DMRG. All results correspond to 100 spins and have  $J_1 = 1$ . We use three GRU layers, where each layer has 100 units. Note that  $J_2 = 0.5$  corresponds to the Majumdar-Ghosh model where the ground state is a product state of spin singlets. For the estimation of the variational energies we use  $4 \times 10^6$  samples.

$J_2$	$E/N$		
	No sign	Marshall sign	DMRG
0.0	-0.4412480(2)	-0.4412760(1)	-0.4412773
0.2	-0.4073635(3)	-0.4073871(3)	-0.4073881
0.5	-0.3749958(6)	-0.3750006(6)	-0.3750000
0.8	-0.4205478(13)	-0.4205695(12)	-0.4207006

RNN wave function with one GRU layer and 50 memory units.

During the completion of this paper, we became aware of another way to estimate entanglement entropies using autoregressive models with conditional sampling [88].

APPENDIX F: TABLES OF RESULTS

In Table I, we state the variational energies of the cRNN wave function for the 1D  $J_1$ - $J_2$  model and compare with results from DMRG. We examine two different methods of training. First, we do not impose an initial sign structure while, secondly, we introduce a background Marshall sign. The results suggest that using a Marshall sign improves the results significantly for  $J_2 = 0.0, 0.2,$  and  $0.5$  (with  $J_1 = 1$  for all cases). We note that our cRNN wave function recovers the sign structure of the ground state if we train it without an initial Marshall sign.

In Table II, we compare the variational energies per site of the 2D TFIM with a lattice size of  $12 \times 12$  for different values of the transverse magnetic field  $h$ , for a 1D pRNN

TABLE II. Variational energies per site for a 1D pRNN wave function (three layers of GRUs with 100 memory units), 2D pRNN wave function (a single layer of 2D vanilla RNN with 100 memory units), PixelCNN wave functions with results taken from Ref. [56], and DMRG (with bond dimension  $\chi = 512$  for  $h = 2$  and  $\chi = 1024$  for both  $h = 3, 4$ ). As a benchmark, we use the 2D TFIM with a lattice size of  $12 \times 12$  for different values of  $h$  where the critical point is at  $h \approx 3$ . Values in bold font correspond to the lowest variational energies and hence to the most accurate estimations of the ground-state energy across all four methods. For the estimation of the variational energy of the trained 1D and 2D pRNN wave functions, we use  $2 \times 10^6$  samples.

$h$	$E/N$			
	1DRNN	2DRNN	PixelCNN	DMRG
2	-2.4 096 018(2)	<b>-2.40 960 262(9)</b>	-2.4 096 022(2)	<b>-2.40 960 263</b>
3	-3.1 738 969(5)	<b>-3.1 739 018(2)</b>	-3.1 739 005(5)	-3.1 738 9966
4	-4.1 217 969(3)	<b>-4.12 179 808(6)</b>	-4.1 217 979(2)	-4.12 179 793

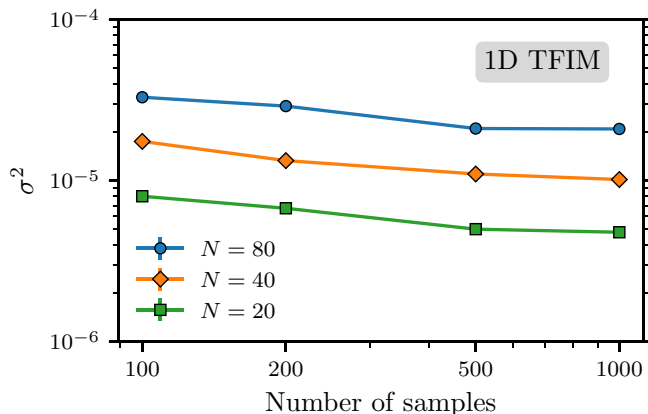


FIG. 9. The energy variance per spin against the number of samples, which suggests that the energy variance saturates and does not improve further by using a larger number of samples for training.

wave function, a 2D pRNN wave function, a PixelCNN wave function [56], and DMRG.

APPENDIX G: SCALING OF RESOURCES (CONTINUED)

Figure 9 shows the dependence of  $\sigma^2$  on the number of samples used to estimate the gradients of the variational energy (see Appendix C). We investigate this effect for the case of the 1D TFIM, using 50 memory units in the pRNN wave function. Even though a large number of samples yields higher statistical accuracy of the gradient estimates used in our optimizations, we observe only a weak dependence of  $\sigma^2$  on the number of samples for all studied system sizes.

In Fig. 10 we present results for the dependence of  $\sigma^2$  on the depth of the pRNN wave function architecture for a critical TFIM with  $N = 40$  sites. We investigate architectures up to a depth of four layers. The number of memory units is adapted such that we have a similar number of variational parameters

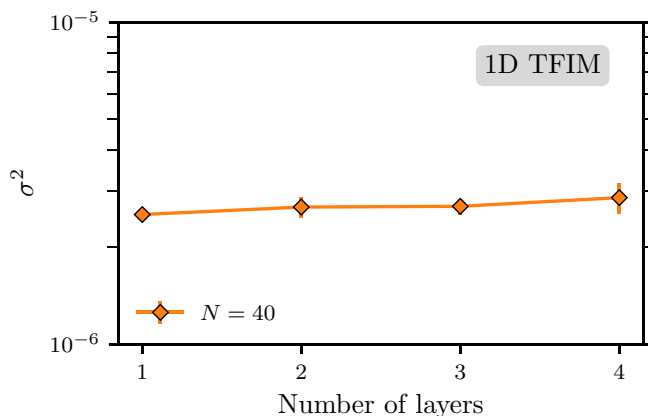


FIG. 10. Scaling study of the energy variance per spin vs the number of layers of a pRNN wave function such that all pRNN wave functions with different layers have approximately the same number of variational parameters. The results show that fixing the number of parameters while changing the number of layers does not affect the energy variance obtained by the pRNN wave function.

TABLE III. Hyperparameters used to obtain the results reported in this paper. Note that the number of samples stands for the batch size used to train the RNN wave function. Multiple seeds are used for the scaling of resources study to provide error bars on our results.

Figures	Hyperparameter	Value
Fig. 3	Architecture	One-layer 1D pRNN wave function with 50 memory units
	Number of samples	$N_s = 1000$ ( $N = 20$ ), $N_s = 500$ ( $N = 80$ ), $N_s = 200$ ( $N = 1000$ )
	Training iterations	20 000
	Learning rate	$5 \times 10^{-3}$
	Seed	111
Fig. 4	Architecture	Three-layer 1D cRNN wave function with 100 memory units
	Number of samples	500
	Training iterations	100 000
	Learning rate	$(\eta^{-1} + 0.1t)^{-1}$ with $\eta = 2.5 \times 10^{-4}$
	Seed	111
Fig. 5(c): 1DRNN	Architecture	Three-layer 1D pRNN wave function with 100 memory units
	Number of samples	500
	Training iterations	150 000
	Learning rate	$(\eta^{-1} + 0.1t)^{-1}$ with $\eta = 10^{-3}$
	Seed	333
Fig. 5(c): 2DRNN	Architecture	One-layer 2D pRNN wave function with 100 memory units
	Number of samples	500
	Training iterations	150 000
	Learning rate	$\eta(1 + t/5000)^{-1}$ with $\eta = 5 \times 10^{-3}$
	Seed	111
Fig. 6(a)	Architecture	One-layer 1D pRNN wave function
	Number of samples	500
	Training iterations	10 000
	Learning rate	$10^{-3}$
	Seeds	111,222,333,444,555
Fig. 6(b)	Architecture	One-layer 1D pRNN wave function
	Number of samples	500
	Training iterations	10 000
	Learning rate	$(\eta^{-1} + 0.1t)^{-1}$ with $\eta = 10^{-3}$
	Seeds	111,222,333,444,555,666,777,888,999,1111
Fig. 9	Architecture	One-layer 1D pRNN wave function with 50 memory units
	Training iterations	10 000
	Learning rate	$10^{-3}$
	Seeds	111,222,333,444,555
	Fig. 10	Architecture
Number of samples		500
Training iterations		10 000
Learning rate		$5 \times 10^{-3}$
Seeds		111,222,333,444,555

( $\approx 31\,000$ ) for each of the four architectures. We find that  $\sigma^2$  depends only weakly on the number of layers.

#### APPENDIX H: HYPERPARAMETERS

In Table III, we present the hyperparameters used to train the RNN wave functions in this paper. We anticipate that

further improvements such as the use of stochastic reconfiguration [13] or a computationally cheaper variant such as K-FAC [71] for the optimization could potentially lead to more accurate estimations of the ground-state energies as compared to the ADAM optimizer [82]. Seeds are listed in the table for reproducibility purposes.

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Neural Information Pro-

cessing Systems Vol. 12 (Curran, New York, 2012), Vol. 1, pp. 1097–1105.

[2] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature (London)* **521**, 436 (2015).

- [3] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
- [4] T. Young, D. Hazarika, S. Poria, and E. Cambria, Recent trends in deep learning based natural language processing [Review Article], *IEEE Comput. Intelligence Mag.* **13**, 55 (2018).
- [5] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, Applications of machine learning in drug discovery and development, *Nat. Rev. Drug Discov.* **18**, 463 (2019).
- [6] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. Paixão, F. Mutz, L. Veronese, T. Oliveira-Santos, and A. F. De Souza, Self-driving cars: A survey, [arXiv:1901.04407](https://arxiv.org/abs/1901.04407) (2019).
- [7] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, Mastering the game of go without human knowledge, *Nature (London)* **550**, 354 (2017).
- [8] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
- [9] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
- [10] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* **136**, B864 (1964).
- [11] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Theory of superconductivity, *Phys. Rev.* **108**, 1175 (1957).
- [12] R. B. Laughlin, Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations, *Phys. Rev. Lett.* **50**, 1395 (1983).
- [13] F. Becca and S. Sorella, *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University, Cambridge, England, 2017).
- [14] R. Orús, Tensor networks for complex quantum systems, *Nat. Rev. Phys.* **1**, 538 (2019).
- [15] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O’Brien, A variational eigenvalue solver on a photonic quantum processor, *Nat. Commun.* **5**, 4213 (2014).
- [16] J. Androsiuk, L. Kulak, and K. Sienicki, Neural network solution of the Schrödinger equation for a two-dimensional harmonic oscillator, *Chem. Phys.* **173**, 377 (1993).
- [17] I. E. Lagaris, A. Likas, and D. I. Fotiadis, Artificial neural network methods in quantum mechanics, *Comput. Phys. Commun.* **104**, 1 (1997).
- [18] M. Sugawara, Numerical solution of the Schrödinger equation by neural network and genetic algorithm, *Comput. Phys. Commun.* **140**, 366 (2001).
- [19] R. G. Melko, G. Carleo, J. Carrasquilla, and J. I. Cirac, Restricted Boltzmann machines in quantum physics, *Nat. Phys.* **15**, 887 (2019).
- [20] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.* **9**, 1735 (1997).
- [21] A. Graves, Supervised sequence labelling, in *Supervised Sequence Labelling with Recurrent Neural Networks* (Springer, New York, 2012), pp. 5–13.
- [22] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Doha, Qatar, 2014), pp. 1724–1734.
- [23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, [arXiv:1412.3555](https://arxiv.org/abs/1412.3555) (2014).
- [24] Z. C. Lipton, J. Berkowitz, and C. Elkan, A critical review of recurrent neural networks for sequence learning, [arXiv:1506.00019](https://arxiv.org/abs/1506.00019) (2015).
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 5998–6008.
- [26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), pp. 4171–4186.
- [27] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, Xlnet: Generalized Autoregressive Pretraining for Language Understanding, *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), pp. 5753–5763.
- [28] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Reconstructing quantum states with generative models, *Nat. Mach. Intell.* **1**, 155 (2019).
- [29] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, Quantum Entanglement in Deep Learning Architectures, *Phys. Rev. Lett.* **122**, 065301 (2019).
- [30] S. Bengio and Y. Bengio, Taking on the curse of dimensionality in joint distributions using neural networks, *IEEE Trans. Neural Netw.* **11**, 550 (2000).
- [31] B. Uribe, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, Neural autoregressive distribution estimation, *J. Mach. Learn. Res.* **17**, 7184 (2016).
- [32] D. Wu, L. Wang, and P. Zhang, Solving Statistical Mechanics Using Variational Autoregressive Networks, *Phys. Rev. Lett.* **122**, 080602 (2019).
- [33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT, Cambridge, MA, 2016).
- [34] I. Goodfellow, Nips 2016 tutorial: Generative adversarial networks, [arXiv:1701.00160](https://arxiv.org/abs/1701.00160) (2016).
- [35] Y. Bengio, P. Simard, and P. Frasconi, Learning long-term dependencies with gradient descent is difficult, *IEEE Trans. Neural Netw.* **5**, 157 (1994).
- [36] J. F. Kolen and S. C. Kremer, Gradient flow in recurrent nets: The difficulty of learning longterm dependencies, in *A Field Guide to Dynamical Recurrent Networks* (IEEE, New York, 2001), pp. 237–243.
- [37] R. Pascanu, T. Mikolov, and Y. Bengio, On the difficulty of training recurrent neural networks, in *Proceedings of the*

- International Conference on Machine Learning (ICML, 2013)*, pp. 1310–1318.
- [38] M. Fannes, B. Nachtergaele, and R. F. Werner, Finitely correlated states on quantum spin chains, *Commun. Math. Phys.* **144**, 443 (1992).
- [39] H. Shen, Mutual information scaling and expressive power of sequence models, [arXiv:1905.04271](https://arxiv.org/abs/1905.04271) (2019).
- [40] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (Association for Computational Linguistics, Doha, Qatar, 2014), pp. 103–111.
- [41] S. Bravyi, D. P. Divincenzo, R. Oliveira, and B. M. Terhal, The complexity of stoquastic local Hamiltonian problems, *Quantum Inf. Comput.* **8**, 361 (2008).
- [42] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, *Nat. Phys.* **14**, 447 (2018).
- [43] S. R. White, Density Matrix Formulation for Quantum Renormalization Groups, *Phys. Rev. Lett.* **69**, 2863 (1992).
- [44] C. Roberts, A. Milsted, M. Ganahl, A. Zalcman, B. Fontaine, Y. Zou, J. Hidary, G. Vidal, and S. Leichenauer, TensorNetwork: A library for physics and machine learning, [arXiv:1905.01330](https://arxiv.org/abs/1905.01330) (2019).
- [45] A. A. Abrikosov, I. Dzyaloshinskii, L. P. Gorkov, and R. A. Silverman, *Methods of Quantum Field Theory in Statistical Physics* (Dover, New York, 1975).
- [46] S. T. Flammia, A. Hamma, T. L. Hughes, and X.-G. Wen, Topological Entanglement Rényi Entropy and Reduced Density Matrix Structure, *Phys. Rev. Lett.* **103**, 261601 (2009).
- [47] S. Ryu and T. Takayanagi, Holographic Derivation of Entanglement Entropy from the Anti-De Sitter Space/Conformal Field Theory Correspondence, *Phys. Rev. Lett.* **96**, 181602 (2006).
- [48] M. B. Hastings, I. González, A. B. Kallin, and R. G. Melko, Measuring Renyi Entanglement Entropy in Quantum Monte Carlo Simulations, *Phys. Rev. Lett.* **104**, 157201 (2010).
- [49] S. R. White and I. Affleck, Dimerization and incommensurate spiral spin correlations in the zigzag spin chain: Analogies to the Kondo lattice, *Phys. Rev. B* **54**, 9862 (1996).
- [50] S. Eggert, Numerical evidence for multiplicative logarithmic corrections from marginal operators, *Phys. Rev. B* **54**, R9612(R) (1996).
- [51] F. Becca, L. Capriotti, A. Parola, and S. Sorella, Variational Wave Functions for Frustrated Magnetic Models, in *Introduction to Frustrated Magnetism: Materials, Experiments, Theory*, edited by C. Lacroix, P. Mendels, and F. Mila (Springer, Berlin, Heidelberg, 2011), pp. 379–406.
- [52] W. Marshall, Antiferromagnetism, *Proc. R. Soc. A* **232**, 48 (1955).
- [53] K. Choo, T. Neupert, and G. Carleo, Two-dimensional frustrated  $j_1 - j_2$  model studied with neural network quantum states, *Phys. Rev. B* **100**, 125124 (2019).
- [54] G. Torlai, J. Carrasquilla, M. T. Fishman, R. G. Melko, and M. P. A. Fisher, Wavefunction positivation via automatic differentiation, [arXiv:1906.04654](https://arxiv.org/abs/1906.04654) (2019).
- [55] J. Thibaut, T. Roscilde, and F. Mezzacapo, Long-range entangled-plaquette states for critical and frustrated quantum systems on a lattice, *Phys. Rev. B* **100**, 155148 (2019).
- [56] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems, *Phys. Rev. Lett.* **124**, 020503 (2020).
- [57] F. Verstraete and J. I. Cirac, Renormalization algorithms for quantum-many body systems in two and higher dimensions, [arXiv:cond-mat/0407066](https://arxiv.org/abs/cond-mat/0407066) (2004).
- [58] S. Yan, David A. Huse, and S. R. White, Spin-liquid ground state of the  $s = 1/2$  kagome heisenberg antiferromagnet, *Science* **332**, 1173 (2011).
- [59] G. Evenbly and G. Vidal, Tensor Network Renormalization, *Phys. Rev. Lett.* **115**, 180405 (2015).
- [60] Anders W. Sandvik, Computational studies of quantum spin systems, in *Lectures on the Physics of Strongly Correlated Systems XIV: Fourteenth Training Course in the Physics of Strongly Correlated Systems*, edited by A. Avella and F. Mancini, AIP Conf. Proc. No. 1297 (AIP, New York, 2010).
- [61] H. W. J. Blöte and Y. Deng, Cluster Monte Carlo simulation of the transverse Ising model, *Phys. Rev. E* **66**, 066110 (2002).
- [62] L. Vanderstraeten, J. Haegeman, P. Corboz, and F. Verstraete, Gradient methods for variational optimization of projected entangled-pair states, *Phys. Rev. B* **94**, 155123 (2016).
- [63] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, Conditional Image Generation with PixelCNN Decoders, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 4790–4798.
- [64] C. Gros, Criterion for a good variational wave function, *Phys. Rev. B* **42**, 6835 (1990).
- [65] R. Assaraf and M. Caffarel, Zero-variance zero-bias principle for observables in quantum monte carlo: Application to forces, *J. Chem. Phys.* **119**, 10536 (2003).
- [66] C. Hubig, J. Haegeman, and U. Schollwöck, Error estimates for extrapolations with matrix-product states, *Phys. Rev. B* **97**, 045125 (2018).
- [67] M. C. Bañuls, D. A. Huse, and J. I. Cirac, Entanglement and its relation to energy variance for local one-dimensional Hamiltonians, *Phys. Rev. B* **101**, 144305 (2020).
- [68] M. Hermans and B. Schrauwen, Training and analysing deep recurrent neural networks, in *Advances in Neural Information Processing Systems*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran, New York, 2013), Vol. 26, pp. 190–198.
- [69] S. Chang, Y. Zhang, W. Han, M. Yu, X. Guo, W. Tan, X. Cui, M. Witbrock, M. Hasegawa-Johnson, and T. S. Huang, Dilated recurrent neural networks, *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 77–87.
- [70] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, [arXiv:1409.0473](https://arxiv.org/abs/1409.0473) (2014).
- [71] J. Martens, J. Ba, and M. Johnson, Kronecker-factored curvature approximations for recurrent neural networks, in *Proceedings of the International Conference on Learning Representations (ICLR, 2018)*.
- [72] K. Choo, A. Mezzacapo, and G. Carleo, Fermionic neural-network states for *ab-initio* electronic structure, *Nat. Commun.* **11**, 2368 (2020).



- [73] J. Carrasquilla, D. Luo, F. Pérez, A. Milsted, B. K. Clark, M. Volkovs, and L. Aolita, Probabilistic simulation of quantum circuits with the transformer, [arXiv:1912.11052](https://arxiv.org/abs/1912.11052) (2019).
- [74] Please see [http://github.com/mhibatallah/RNNWavefunctions](https://github.com/mhibatallah/RNNWavefunctions).
- [75] C. Roth, Iterative retraining of quantum spin models using recurrent neural networks, [arXiv:2003.06228](https://arxiv.org/abs/2003.06228) (2020).
- [76] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TENSORFLOW: Large-scale machine learning on heterogeneous systems, 2015, software available from [tensorflow.org](https://tensorflow.org).
- [77] J. Appleyard, T. Kocisky, and P. Blunsom, Optimizing performance of recurrent neural networks on GPUs, [arXiv:1604.01946](https://arxiv.org/abs/1604.01946) (2016).
- [78] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, Pixel Recurrent Neural Networks, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, Vol. 48 (JMLR, New York, 2016), pp. 1747–1756.
- [79] Y. Wu, S. Zhang, Y. Zhang, Y. Bengio, and R. R. Salakhutdinov, On multiplicative integration with recurrent neural networks, in *Proceedings of the Advances in Neural Information Processing Systems* (NeurIPS, 2016), pp. 2856–2864.
- [80] A. W. Sandvik and G. Vidal, Variational Quantum Monte Carlo Simulations with Tensor-Network States, *Phys. Rev. Lett.* **99**, 220602 (2007).
- [81] S.-X. Zhang, Z.-Q. Wan, and H. Yao, Automatic differentiable Monte Carlo: Theory and application, [arXiv:1911.09117](https://arxiv.org/abs/1911.09117) (2019).
- [82] D. P. Kingma and J. Ba, ADAM: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
- [83] R. Assaraf and M. Caffarel, Zero-Variance Principle for Monte Carlo Algorithms, *Phys. Rev. Lett.* **83**, 4682 (1999).
- [84] B. Clark, Variational Monte Carlo notes for boulder summer school 2010 (unpublished).
- [85] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, Monte Carlo gradient estimation in machine learning, [arXiv:1906.10652](https://arxiv.org/abs/1906.10652) (2019).
- [86] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, in *Proceedings of Advances in Neural Information Processing Systems* (NeurIPS, 2000), pp. 1057–1063.
- [87] E. Lieb and D. Mattis, Ordering energy levels of interacting spin systems, *J. Math. Phys.* **3**, 749 (1962).
- [88] Z. Wang and E. J. Davis, Calculating Rényi entropies with neural autoregressive quantum states, in *Proceedings of the Machine Learning and Physical Sciences Workshop* (NeurIPS, 2019), [arXiv:2003.01358](https://arxiv.org/abs/2003.01358).