

# Transformer Wave Function for Quantum Long-Range models

Sebastián Roca-Jerat,<sup>1,2</sup> Manuel Gallego,<sup>1,3</sup> Fernando Luis,<sup>1,2</sup> Jesús Carrete,<sup>1,2</sup> and David Zueco<sup>1,2,\*</sup>

<sup>1</sup>*Instituto de Nanociencia y Materiales de Aragón (INMA),  
CSIC-Universidad de Zaragoza, Zaragoza 50009, Spain*

<sup>2</sup>*Departamento de Física de la Materia Condensada,  
Universidad de Zaragoza, Zaragoza 50009, Spain*

<sup>3</sup>*Departamento de Física Teórica, Universidad de Zaragoza, Zaragoza 50009, Spain*

(Dated: January 8, 2025)

We employ a neural-network architecture based on the Vision Transformer (ViT) architecture to find the ground states of quantum long-range models, specifically the transverse-field Ising model for spin-1/2 chains across different interaction regimes. Harnessing the transformer’s capacity to capture long-range correlations, we compute the full phase diagram and critical properties of the model, in both the ferromagnetic and antiferromagnetic cases. Our findings show that the ViT maintains high accuracy across the full phase diagram. We compare these results with previous numerical studies in the literature and, in particular, show that the ViT has a superior performance than a restricted-Boltzmann-machine-like ansatz.

## I. INTRODUCTION

Understanding quantum many-body problems is crucial for elucidating the properties of condensed matter systems. However, this is generally a challenging task, as exemplified by the problem of searching for the ground states of strongly correlated models. Beyond exactly solvable models, some systems are tractable by introducing elements such as matrix product states (MPS), tensor networks [1], dynamical mean field theory (DMFT) [2], quantum Monte Carlo [3], and hybrid quantum-classical algorithms [4]. Additionally, there are physically inspired ansätze available, like the Bethe and polaron approaches, correlator product states [5], and generalized coherent states [6].

In 2017, neural networks (NNs) emerged as an alternative ansatz for solving quantum many-body systems [7]. NNs have since been recognized for their ability to generate states beyond the area law entanglement [8, 9], addressing problems beyond the reach of MPS. Variational optimization of ansätze like NNs also avoids the sign problem of other quantum Monte Carlo techniques, enabling the handling of frustrated systems and account for states with long correlation lengths beyond those tractable with DMFT. Despite these advantages, it is difficult to determine in advance which NN architecture will be most suitable for a specific problem or whether it can be trained with sufficient precision. Therefore, it is interesting to investigate generic models of significant importance and evaluate how well they are solved using NNs [10].

In this context, we focus on a set of models in which two-body interactions decay according to a power law, commonly referred to as long-range systems. They are ubiquitous in nature, with examples arising, among others, from dipolar [11], Coulomb [12], and van der Waals interactions [13]. Recent experimental advances

in atomic, molecular, and optical systems have renewed interest in long-range models [14–17]. In these experiments, the effective interactions are often long ranged and tunable, underscoring the need for a comprehensive understanding of long-range systems. Although less studied than their short-ranged counterparts, there are already some rigorous and numerical results available for systems with long-range interactions [18–22]. Various equilibrium and dynamical properties have been explored in comparison with short-ranged systems. Notable examples include the existence (or absence) of an area law of entanglement [23–26], the algebraic decay of two-point correlators outside criticality [27–29], the spreading of correlations [30], the existence of Majorana modes [31], and topological properties [32].

In this work, we test neural networks (NNs) for solving quantum long-range models, specifically focusing on the quantum Ising model as a paradigmatic example. Our primary motivation is to use a sufficiently generic NN ansatz capable of covering the entire phase diagram, from short-range to strong long-range interactions, and for both ferromagnetic and antiferromagnetic phases. Since the introduction of Restricted Boltzmann Machines (RBMs), several NN architectures have been developed, including multilayer perceptrons [33], convolutional neural networks [34], and recurrent architectures [35]. This list is not exhaustive; for a comprehensive review of the field, we refer to Ref. 10. Recently, the transformer [36] has proven to be a game changer in the field of machine learning (ML), replacing recurrent architectures in natural language processing tasks [37], being adapted with particular success adapted to computer vision tasks in the form of the Vision Transformer (ViT) [38] but also to more specialized applications like the determination of protein structures [39]. A key feature of transformer architectures is their decoupling of nonlinear processing, which may still be performed by conventional fully connected subnetworks acting on subsets of the data, and mixing between those pieces of data according to the relevance of those interactions to the result, which is han-

---

\* [dzueco@unizar.es](mailto:dzueco@unizar.es)

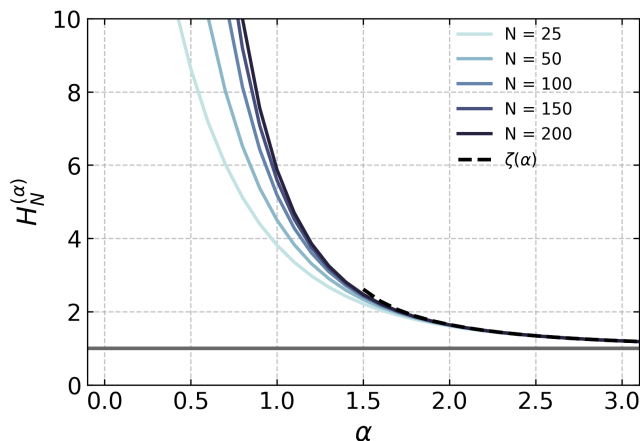


FIG. 1. Dependence of the generalised harmonic numbers on the system’s size ( $N$ ) and on the range of the interactions ( $\alpha$ ). It can be seen that for  $\alpha < 1.5$  the dependence of this factor on size intensifies. The solid grey line indicates the asymptotic value of  $H_N^{(\alpha)} = 1$  and the dashed black line indicates the Riemann zeta function,  $\zeta(\alpha)$ , to which the generalised harmonic number tends when  $N \rightarrow \infty$ .

dled by an attention mechanism. This allows it to extract complex patterns very efficiently. Due to these promising features, application of transformers in the field of condensed matter as a variational ansatz has quickly followed [40–42]. This type of architecture could help finding solutions more efficiently in problems where established methods like MPS (systems with more than one dimension) or quantum Monte Carlo encounter difficulties (wave function with negative coefficients).

In this work, we continue to explore the capabilities of this type of architecture by selecting the Vision Transformer (ViT). We demonstrate that this architecture is capable of computing the full phase diagram and characterizing its critical properties. By doing so, we provide a single ansatz that encompasses previous numerical works discussing parts of the phase diagram [43–47]. Additionally, we compare the performance of our ViT and RBM architectures. Our results help establish transformers as an efficient variational ansatz that can be highly useful for tackling a wide range of problems in many-body physics.

The rest of the manuscript is organized as follows: in section II we lay out the details of the family of models we are trying to solve, in section III we discuss the methods, software and parameters that we use, in section IV we present and contextualize our results, and in the last section we summarize our main conclusions. Details of the finite size scaling are sent to Appendix A.

## II. LONG-RANGE INTERACTING QUANTUM MODELS

Generically, quantum long-range models on a  $d$ -dimensional,  $N$ -site lattice contain interactions described by a Hamiltonian of the form

$$\mathcal{H}_c = -\frac{1}{2} \sum_{ij} J_{ij} (\mathcal{C}_i \mathcal{C}_j^\dagger + h.c.), \quad (1)$$

where  $\mathcal{C}_i$  is a local operator acting on site  $i$  and the interactions decay according to a power law

$$J_{ij} = J \frac{\tilde{J}_{ij}}{\tilde{N}}, \quad (2)$$

with

$$\tilde{J}_{ij} = \begin{cases} b & \text{if } i = j \\ r_{ij}^{-\alpha} & \text{otherwise.} \end{cases} \quad (3)$$

In this equation,  $b$  is a parameter that can be tuned to shift the spectrum of  $J$  and is taken as  $b = 1$  in our simulations, and the parameter  $\alpha$  sets the range of the interactions. Throughout this paper, we will assume periodic boundary conditions (PBC) and  $d = 1$ , therefore we define  $r_{ij} = \min(|i - j|, N - |i - j|)$ .

It is important to discuss the normalization used in Eq. (2). For  $\alpha < d$ , the interactions decay slowly enough that the sum in the coupling term (1) depends on  $N$  superlinearly, breaking the extensive character of the model [19, 21, 48]. Kac’s renormalization factor,  $1/\tilde{N}$ , restores that feature, ensuring a well-defined thermodynamic limit. In particular,

$$\tilde{N} = \sum_j \tilde{J}_{ij} = 1 + \sum_{j=1}^N \frac{1}{r_{ij}^\alpha} = 1 + H_{\lfloor N/2 \rfloor}^{(\alpha)} + H_{N-1-\lfloor N/2 \rfloor}^{(\alpha)}, \quad (4)$$

where  $H_N^{(\alpha)}$  is a generalized harmonic number. We have used the fact PBC make the model translationally invariant and thus  $\sum_i \tilde{J}_{ij}$  is independent of  $j$ . In figure 1, we plot  $H_N^{(\alpha)}$  for different sizes as a function of  $\alpha$ . For  $\alpha > 1$ ,  $H_N^{(\alpha)}$  approaches the  $\zeta(\alpha)$  function as  $N \rightarrow \infty$ . Indeed, for  $\alpha \gtrsim 2$ , in the next sections we confirm through numerical simulations that the dependence of the Kac factor on system size can be safely neglected. Eventually,  $\tilde{N} \rightarrow 3$  as  $\alpha \rightarrow \infty$ . For  $\alpha < 1$ , the asymptotics of  $H_N^{(\alpha)}$  are richer, but it diverges as a function of  $N$ , which makes this regularization necessary to have a well-defined thermodynamic limit.

The general phenomenology, both quantum and classical, can be characterized by the value of  $\alpha$ . The non-additive regime, characterized by  $\alpha < 1$ , is termed (*strong*) *long-range*. This regime is often overlooked in many analytical and numerical studies due to the challenges posed by an ill-defined thermodynamic limit resulting from non-extensivity. However, Kac’s rescaling

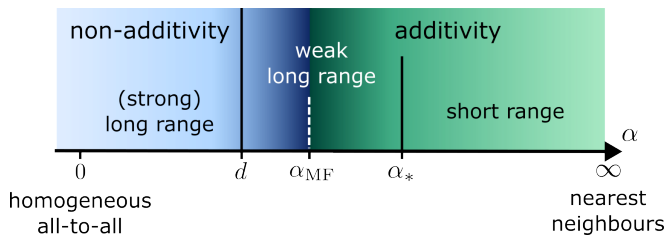


FIG. 2. Diagram of the nature of interactions as a function of  $\alpha$ .

overcomes this obstacle. In the quantum context, numerical studies that solve the transverse-field Ising model in the strong long-range regime can be found in Refs. 43 and 49, confirming that in this regime, the model is within the mean-field universality class. Furthermore, it has been proven that, in both the classical and quantum realms, some models can be solved exactly through mean-field-like approximations that become exact in this regime [45, 50–54].

Within  $\alpha > 1$  domain, additional subregimes are discernible: a threshold  $\alpha_*$  exists such that for  $\alpha > \alpha_*$  the critical exponents of the model align with those of the nearest-neighbour model ( $\alpha \rightarrow \infty$ ). This is the *short-range regime*. Conversely, in the range where  $1 < \alpha < \alpha_*$ , the model exhibits critical exponents distinct from their short-range counterparts, revealing the influence of long-range interactions while retaining additivity. This condition defines the *weak long-range regime* [18, 21]. It is also possible to further subdivide these classes by introducing  $\alpha_{MF}$  such that for  $1 < \alpha < \alpha_{MF}$  the critical exponents correspond to those predicted by mean-field (MF) theories [55, 56]. It appears that the weak long-range regime is the most challenging to access from an analytical standpoint, making numerical approaches indispensable for its understanding [23, 43, 46, 47, 57]. For ease of reference, we summarize this classification in Fig. 2.

In this paper, we solve the long-range Ising model, which is of significant interest for various experiments [14, 58, 59] and from a theoretical perspective, offering numerous results for comparison. We consider the Hamiltonian given by

$$\mathcal{H} = \sum_{ij}^N J_{ij} \sigma_i^z \sigma_j^z - h_x \sum_i^N \sigma_i^x, \quad (5)$$

where the  $\sigma^x$  represent the standard Pauli matrices and  $J_{ij}$  is defined as in Eq. (2). Throughout all computations, we set  $h_x = 1$ , thereby fixing the units of  $J$ . This model yields exact analytical results in the limit  $\alpha \rightarrow \infty$  (nearest neighbors) via the Jordan-Wigner transformation. Conversely, in the limit  $\alpha \rightarrow 0$ , the model reduces to the Lipkin-Meshkov-Glick (LMG) model, which is also solvable in the thermodynamic limit. Moreover, for  $\alpha < d$ , the model admits an exact solution in the ferromagnetic ( $J < 0$ ) case [45]. These analytical insights

can be complemented by numerical approaches that span various ranges of  $\alpha$ , utilizing diverse techniques such as quantum Monte Carlo [43, 49] and ansätze such as tensor networks [23, 46], and, more recently, restricted Boltzmann machines (RBM) [47]. All these works together offer a comprehensive overview for both the ferromagnetic and antiferromagnetic scenarios. The system undergoes a second-order quantum phase transition (QPT) from ferromagnetic and Néel respectively ordered phases to paramagnetic phases at a certain critical point  $J_c(\alpha)$ . Within this model,  $\alpha_{MF} = 5/3$ , below which the critical exponents align with those predicted by Mean Field theory. Conversely, for  $\alpha > 3$ , the system falls within the universality class of the 2D classical short-range Ising model. As a consequence, model (5) provides an excellent baseline for evaluating the accuracy of the transformer variational ansatz across the entire  $\alpha$  spectrum, ranging from short-range to flat interactions, and for both ferromagnetic and antiferromagnetic cases.

### III. METHODS

#### A. Variational quantum Monte Carlo

We use an unnormalized representation of the ground state over the canonical basis states  $\{|s\rangle\}$ :

$$|\psi_{\theta}\rangle = \sum_{\mathbf{s}} \psi_{\theta}(\mathbf{x}_{\text{in}}) |s\rangle, \quad (6)$$

where  $\psi_{\theta}$  is a sufficiently flexible ansatz with a collection of parameters  $\theta$ , yielding the probability amplitude of each of those states, and  $\mathbf{x}_{\text{in}}$  is an appropriate encoding of  $|s\rangle$  to serve as an input to the function. In this work we focus on spin 1/2 chains, so  $\mathbf{x}_{\text{in}}$  will be composed of a binary variable for each spin, whose values we encode as +1 or -1 (see Figure 3a). Our choices for  $\psi_{\theta}$  are all general machine-learning models, which we generically represent as  $M(\theta)$ .

In this framework, the solution for the ground state of the system comes in the form of an optimal set of parameters  $\theta$ . Those parameters are thus variationally optimized by minimizing the energy of the Hamiltonian,  $\langle \mathcal{H} \rangle_{\psi_{\theta}} \equiv \langle \psi_{\theta} | \mathcal{H} | \psi_{\theta} \rangle / \langle \psi_{\theta} | \psi_{\theta} \rangle$ . The expected value of an observable explicitly represented on this basis as a function  $Q(\mathbf{x}_{\text{in}}) = \frac{\langle \mathbf{s} | Q | \psi_{\theta} \rangle}{\langle \mathbf{s} | \psi_{\theta} \rangle}$ , can be calculated as

$$\langle Q \rangle_{\psi_{\theta}} = \sum_{\mathbf{s}} p_{\theta}(\mathbf{s}) Q(\mathbf{x}_{\text{in}}), \quad (7)$$

where  $p_{\theta}(\mathbf{s}) = |\psi_{\theta}(\mathbf{x}_{\text{in}})|^2 / \langle \psi_{\theta} | \psi_{\theta} \rangle$ . Note that, in contrast to a more conventional ML workflow for building a classifier or a regressor, no training outputs are necessary here to define a loss, since optimization is guided by the minimization of the value of the energy  $\langle \mathcal{H} \rangle_{\psi_{\theta}}$  estimated using the model itself. The data used to train the model

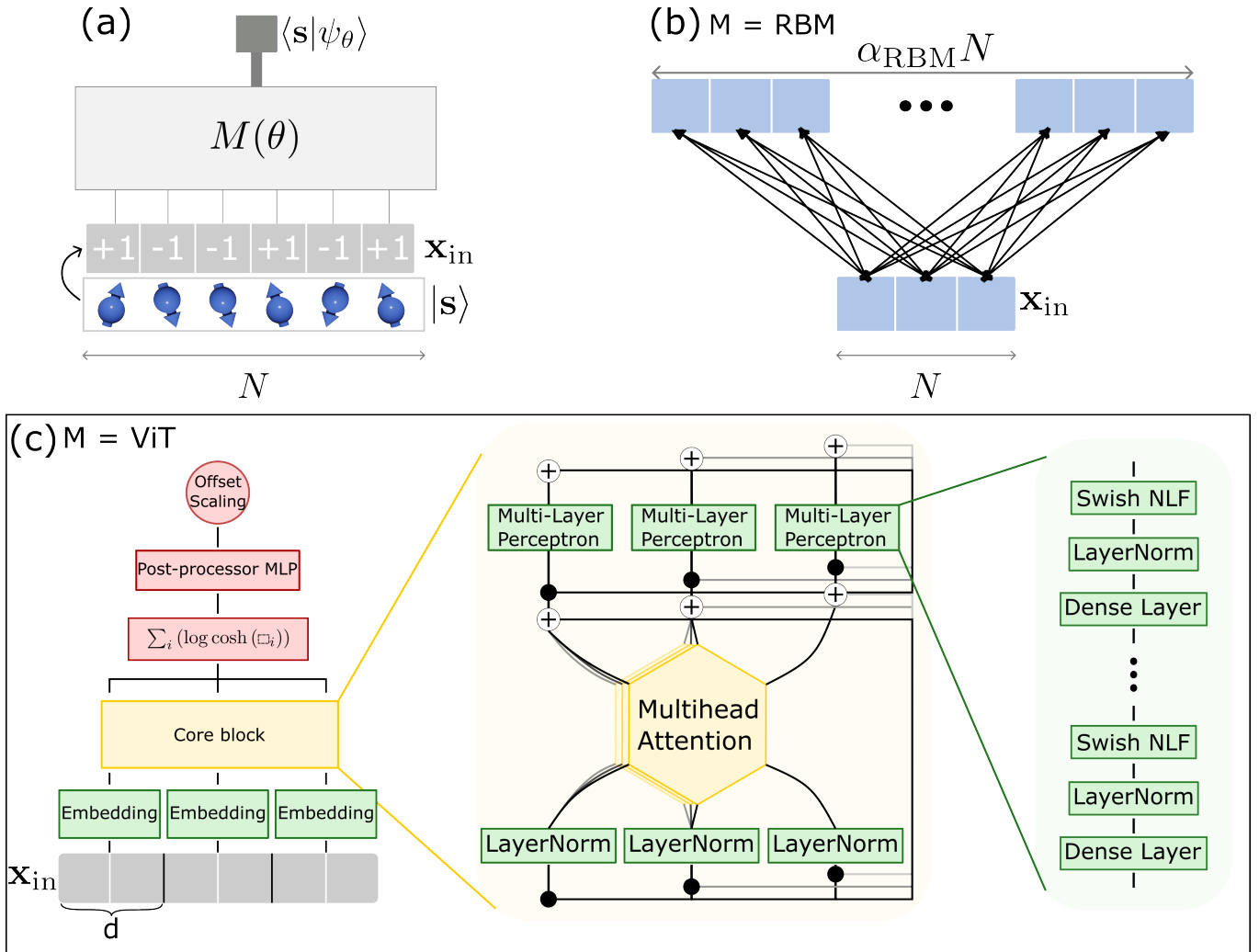


FIG. 3. (a) General scheme of the parameterization of a wave-function,  $\psi_\theta$ , through a machine learning model,  $M$ , characterized by a series of variational parameters  $\theta$ . (b) General structure of a RBM. The parameter  $\alpha_{\text{RBM}}$  indicates the density of the hidden layer. (c) Block diagram representing the structure of the ViT ansatz used in this work. The color code is used: green corresponds to all those operations that act at token level; in yellow we represent operations that generate correlations between the channels originating from different tokens; finally, red highlights the operations placed after the Core block that merge the channels and post-process that result.

are thus only inputs in the form of states sampled from the canonical basis.

As these definitions suggest, we are forced to resort to Monte Carlo (MC) sampling in order to evaluate the averages to carry out the minimization process, given the impossibility of considering the totality of states. Specifically, we use the Metropolis-Hastings algorithm with a combined transition rule. On the one hand, a random spin is flipped according to a uniform probability without taking into account any constraints such as the total magnetization. On the other hand, we invert the total magnetization of a random number of Markov chains. The probability for the former is three times larger than the probability for the latter in our implementation. Furthermore, among all the available optimization methods, we employ Stochastic Gradient Descent (SGD) with cus-

tom schedules for the learning rate,  $\lambda$ , combined with the Stochastic Reconfiguration (SR) method [60], characterized by a stabilizing parameter called diagonal shift that we denote as  $\Delta_{sr}$ .

## B. Ansätze: neural networks and restricted Boltzmann machines

In this variational framework, NNs are leveraged to provide a trainable  $\psi_\theta$ . In this study we employ and compare two different architectures with very different levels of complexity: an RBM adapted for regression and a vision transformer. Multilayer perceptrons (MLPs), the most basic incarnation of a feed-forward neural network (FFNN), are used as building blocks in the ViT and

also serve to introduce this RBM.

The architecture of an FFNN is characterized by the unidirectional flow of information from the input towards the output, passing through a set of hidden layers. In the case of an MLP, each neuron in a layer computes a linear combination of all the outputs from the previous layer and applies a nonlinear activation function to the results to create its own output. The defining characteristics of an MLP are the all-to-all connectivity between adjacent layers and the fact that there is at least a single hidden layer. The functional form of an MLP with activation function  $F$  and  $\ell$  hidden layers is

$$\begin{aligned} \text{MLP}(\mathbf{x}_{\text{in}}; \mathbf{w}_1, \mathbf{b}_1, \mathbf{w}_2, \mathbf{b}_2 \dots \mathbf{w}_{\text{out}}, \mathbf{b}_{\text{out}}) = \\ = F(\mathbf{b}_{\text{out}} + \mathbf{w}_{\text{out}} F(\mathbf{b}_\ell + \mathbf{w}_\ell F(\dots F(\mathbf{b}_1 + \mathbf{w}_1 \mathbf{x}_{\text{in}})))) \end{aligned}$$

Here, the  $\mathbf{w}$  matrices and  $\mathbf{b}$  vectors are the weights and biases of each layer, respectively. Layers can have different widths, and  $F$  is understood to act elementwise on its arguments. We create the functional form of our ansatz by using an MLP for the logarithm of the wave function:

$$\psi_{\boldsymbol{\theta}}^{\text{FFNN}} = \exp \left[ \sum \text{MLP}(\mathbf{x}_{\text{in}}; \boldsymbol{\theta}) \right]. \quad (8)$$

where  $\boldsymbol{\theta}$  stands for the collection of weights and biases and the sums run over elements of the output to generate a single value. This results in a real wave function with positive coefficients, which aligns with our interests. However, if one were to deal with problems where the ground state is a non-positive wave function, the ansatz would need to be modified. One option can be to add a second MLP to parametrize the imaginary part of the logarithm, i.e., the phase of the wave function. This covers both the scenario where the wave function may be complex and the case where it is real but non-positive. For the latter, one can modulate the imaginary part with a sigmoid function with a maximum amplitude of  $\pi$  to restrict the phase to the values 0 or  $\pi$ .

A perceptron with a single hidden layer and some minor changes is equivalent to a restricted Boltzmann machine (RBM; see Figure 3b), originally devised as a generative model. Specifically, the nonlinear activation function must be chosen as  $F(x) = \log \cosh(x)$  and an additional linear term is added to the output. The ansatz in this case is

$$\begin{aligned} \log \psi_{\boldsymbol{\theta}}^{\text{RBM}} = \mathbf{a} \mathbf{x}_{\text{in}} + \sum_j \log \left[ \cosh((\text{Re } \mathbf{w}) \mathbf{x}_{\text{in}} + \text{Re } \mathbf{b})_j \right] + \\ + i \sum_j \log \left[ \cosh((\text{Im } \mathbf{w}) \mathbf{x}_{\text{in}} + \text{Im } \mathbf{b})_j \right] \end{aligned} \quad (9)$$

In the expression above, the product runs over outputs. The width of the output layer is usually chosen as proportional to the width of the input layer through an integer factor known as the density [7]. This variational ansatz

results in a wave function with parameters  $\boldsymbol{\theta} = \{\mathbf{a}, \mathbf{w}, \mathbf{b}\}$ . We take  $\mathbf{a} = 0$  following the implementation in NETKET [61].

Although on paper MLPs can be used to approximate any sufficiently well-behaved function according to a number of universality theorems [62, 63], the basic MLP architecture faces important limitations in practice arising, among other factors, from the phenomenon of vanishing gradients during training [64] as well as the complexity of their all-to-all mapping. The former has been alleviated through breakthroughs such as normalization layers (e.g. BatchNorm [65] and LayerNorm [66]), residual learning by means of additive shortcuts around groups of layers [67, 68] and new activation functions like Swish [69]. In terms of reduction in complexity, the transformer architecture, originally developed to replace recurrent architectures in machine translation tasks [36], has found applications in all areas of machine learning, and in particular been very successfully adapted to computer vision tasks as the vision transformer (ViT) [38]. A key feature of transformer architectures is their decoupling of nonlinear processing, which is still performed by modified MLPs acting on subsets of the data, and mixing between those pieces of data according to the relevance of those interactions to the result, which is handled by an attention mechanism. The details of the version of the transformer used in this work are presented in the form of a block diagram in Figure 3c, and the main operations are described in what follows.

First of all, *tokenization* takes place. Here the quantum states  $|\mathbf{s}\rangle$ , represented as their corresponding arrays  $\mathbf{x}_{\text{in}}$ , are divided into  $n$ -tokens of dimension  $d$  ( $N = dn$ ). Next they are passed through a linear embedding layer with trainable parameters (common to all tokens), mapping each token into a vector of arbitrary dimension  $d_{\text{emb}}$ . The next stage is an attention block where correlations between tokens are taken into account. As depicted in the central panel of Figure 3c, we first normalize the output of the previous embedding using a LayerNorm operation [66]. We then split each token into  $h$  vectors of dimension  $p = d_{\text{emb}}/h$ . Each of these vectors goes into one of the  $h$  heads of the multihead attention layer, where the following operation takes place:

$$\mathbf{A}_i^\mu = \sum_{j=1}^n a_{ij}^\mu V^\mu \mathbf{x}_j^\mu. \quad (10)$$

Here,  $\mathbf{x}_j^\mu$  represents the  $p$ -dimensional vector processed by the  $\mu$ -th head coming from token  $j$ . As it can be seen, tokens are mixed within each head through the linear transformation parameterized by  $V^\mu$  and the attention matrix  $a_{ij}^\mu$ . Note that this is a simplified version of the original dot-product attention mechanism [36]. The key feature of this simplification is that the attention weights are dependent only on the relative positions of each pair of spins. In contrast, the usual dot-product attention is invariant with respect to permutations and therefore does not encode positions at all, so explicit positional encod-

ings must be introduced to avoid losing that information [70]. Furthermore, since we deal with spin chains with periodic boundary conditions, we enforce the translational symmetry by constraining  $a_{ij}^\mu$  to be a circulant matrix, i.e.,  $a_{ij}^\mu = c_{(j-i) \bmod p}^\mu$  [40]. This is complemented by an explicit symmetrization through an average over all the possible cyclic permutations of the tokens. In this regard, the ViT improves upon the complexity of a fully-connected model when applied to a translationally symmetric system, as the latter would require a similar symmetrization over all cyclic permutations of spins. After applying this layer, we concatenate the  $h$   $\mathbf{A}_i^\mu$  vectors to recover a single  $d_{\text{emb}}$ -dimensional vector for each of the input tokens. Each of those vectors is then fed through a modified MLP for nonlinear processing, whose components are represented in the right-hand-side panel of Figure 3c. The output is again a  $d_{\text{emb}}$ -dimensional vector for each input token. Note that both operations (mixing by the attention block and processing by the MLP) are also surrounded by ResNet-like bypasses [68]. We further pass the output through an elementwise  $\log \cosh(x)$ , which we have found to greatly enhance the training performance of the model when it must learn wave functions whose norm is concentrated over a few states. The inspiration for our choice of transformation comes from the simpler RBM-like architecture, where the log-cosh function is essential for a satisfactory training, as opposed to more conventional nonlinear activation functions. This completes our ViT core block and the set of operations defined terms of tokens. A pooling operation is now applied so that the  $n$  vectors corresponding to the tokens are averaged to give rise to a single (still  $d_{\text{emb}}$ -dimensional) vector. This is further processed by another MLP where layer by layer the dimensions of the vector resulting from pooling are reduced:  $d_{\text{emb}} \rightarrow d_1 \rightarrow \dots \rightarrow d_{n_{pp}}$ . Finally, a last linear transformation acts as an offset-and-scaling block. This complete ViT is used to create the real parts of  $\log \psi_\theta$ , and we do not include an imaginary part.

### C. Parameters and computational details

All calculations in the following studies were performed using GPUs. The hyperparameters used in the ViT architecture are listed in Table I. Our code is written in PYTHON and available on GitHub [71]. Our models are defined using the FLAX neural-network library [72], which runs on top of JAX [73]. Training (including the schedules for the learning rate and for the diagonal shift in the SR preconditioner) is handled by a combination of OPTAX [74] and NETKET [61], which also takes care of sampling.

Each iteration of the training is performed using a total of 4096 MC samples distributed over 1024 independent Markov chains. For the training process we chose a protocol for the learning rate,  $\lambda$ , consisting of a linear warm-up followed by an exponential decay. This protocol is defined by an initial value of the learning rate,  $\lambda_0$ ,

Variable	Notation	Value
Chain length	$N$	50 – 200
Token dimension	$d$	$N/10$
Number of tokens	$n = N/d$	10
Embedding dimension	$d_{\text{emb}}$	14
Number of heads	$h$	2
Head dimension	$p = d_{\text{emb}}/h$	7
MLP layers within Core Block	$n_{\text{MLP}}$	3
Post-processor MLP	$(d_1, \dots, d_{n_{pp}})$	(5,)

TABLE I. Summary of parameters in the ViT architecture and their values used for the simulations [cf. Figure 3].

a maximum value,  $\lambda_{\text{max}}$  that it attains after  $n_{\text{warm}}$  iterations corresponding to the linear warm-up and a ratio  $\gamma$  for the exponential decay. To obtain the states with which we created Figure 4, a maximum of 250 iterations for training were used, along with following parameters:  $\lambda_0 = 0.1$ ,  $\lambda_{\text{max}} = 2.0$ ,  $n_{\text{warm}} = 75$ ,  $\gamma = 0.995$ . The stabiliser parameter present in the SR method follows a simple linear schedule with an initial value of  $\Delta_{sr} = 10^{-2}$  and a final value of  $\Delta_{sr} = 10^{-4}$ .

## IV. RESULTS

Here we discuss the performance of the ViT ansatz for characterizing the full phase diagram of Eq. (5). We compare these results with previous studies that cover specific ranges of the model.

### A. Quantities to compute

We define the *generalized* staggered magnetization through the operator  $m_s$ ,

$$\hat{m}_s(q) = \frac{1}{N} \sum_j \hat{\sigma}_j^z e^{iqj}, \quad (11)$$

where  $q = 0$  when  $J < 0$  and  $q = \pi$  when  $J > 0$ . Once the network parameters have been determined, this can be computed using Eq. (7).

To characterize the entanglement between two partitions of the system within the ground state, we also compute the Renyi-2 entanglement entropy  $\hat{S}_2$  following Ref. 75. We define  $|\psi_\theta\rangle$  based on the states of the computational basis corresponding to each of the partitions,  $\mathbf{s}_A \in A$  and  $\mathbf{s}_{\bar{A}} \in \bar{A}$ , such that  $|\psi_\theta\rangle = \sum_{\mathbf{s}_A, \mathbf{s}_{\bar{A}}} \psi_\theta(\mathbf{s}_A, \mathbf{s}_{\bar{A}}) |\mathbf{s}_A\rangle |\mathbf{s}_{\bar{A}}\rangle$ . Therefore we can define

$$\hat{S}_2 = -\log_2 \text{Tr}[\rho_A^2], \quad (12)$$

where  $\rho_A = \text{Tr}_{\bar{A}} \rho$ . From this definition we can sample states from each of the partitions to estimate  $\langle \hat{S}_2 \rangle$

according to

$$\langle \hat{S}_2 \rangle = \sum_{\substack{\mathbf{s}_A, \mathbf{s}_{\bar{A}} \\ \mathbf{s}'_A, \mathbf{s}'_{\bar{A}}} p_\theta(\mathbf{s}_A, \mathbf{s}_{\bar{A}}) p_\theta(\mathbf{s}'_A, \mathbf{s}'_{\bar{A}}) S_2(\mathbf{s}_A, \mathbf{s}_{\bar{A}}, \mathbf{s}'_A, \mathbf{s}'_{\bar{A}}), \quad (13)$$

where

$$S_2(\mathbf{s}_A, \mathbf{s}_{\bar{A}}, \mathbf{s}'_A, \mathbf{s}'_{\bar{A}}) = -\log \left\langle \frac{\psi_\theta(\mathbf{s}_A, \mathbf{s}'_{\bar{A}}) \psi_\theta(\mathbf{s}'_A, \mathbf{s}_{\bar{A}})}{\psi_\theta(\mathbf{s}_A, \mathbf{s}_{\bar{A}}) \psi_\theta(\mathbf{s}'_A, \mathbf{s}'_{\bar{A}})} \right\rangle. \quad (14)$$

For our simulations we choose  $A$  to be half of the system size. This procedure is already implemented in NetKet [61].

Finally, it is useful to employ a metric to quantify the quality of the converged wave functions. One such metric is the V-score [76]. It is defined in terms of the energy fluctuations, conveniently normalized to yield an  $N$ -independent quantity:

$$\text{V-score} = N \left( \frac{\langle \mathcal{H}^2 \rangle_{\psi_\theta} - \langle \mathcal{H} \rangle_{\psi_\theta}^2}{\langle \mathcal{H} \rangle_{\psi_\theta}^2} \right), \quad (15)$$

## B. Phase diagram and critical exponents

In Figure 4 we summarize our calculations for the full phase diagram of the Ising model (5). In contour 4a) we plot the fluctuations  $\langle \hat{m}_s^2 \rangle$  in order to resolve the phase transition. Here the simulations are done with  $N = 50$  spins and a discretization of  $\delta J = 0.375$  and  $\delta \alpha = 0.4$ . The darker zones stand for ordered phases, both ferro and antiferromagnetic. We also plot several critical  $J_c$  found with different techniques in previous works [43–46]. Light blue circles mark values where a deeper investigation has been done (see below). In panels 4 b) and c), the V-score is plotted for different sizes  $N = 20, 50, 100$  along the cuts marked in panel a) with white lines (the darker the curve, the larger the size). We observe that the V-score does not depend on the system size, demonstrating that the simulations do not deteriorate as  $N$  increases. The values around  $10^{-4}$  compare well with variational methods for other quantum many-body problems [76]. Beyond this figure of merit, in panels 4 d) and e) we plot the fluctuations  $\langle \hat{m}_s^2 \rangle$  for the same sizes and along the same cuts. In addition to the finite-size behavior, the curves are smooth, which confirms the convergence of the method. It is important to note that all of these results are the outcome of single shot runs. No averaging or filtering has been performed, which again highlights the consistent convergence of the model. It is worth stressing that the inclusion of the log-cosh function at the end of the core block [cf. Figure 3], as described in Section III B, is instrumental in obtaining accurate states in the deep FM regime.

Given that the model is able to qualitatively obtain the phase diagram across the entire parameter space, we proceed to extract the critical exponents. This requires

precision in the order parameter that characterizes the transition. In our case, we take advantage of how well the ViT reproduces the fluctuations  $\langle \hat{m}_s^2 \rangle$ . This allows us to determine both the critical point  $J_c$  and the critical exponents associated with the phase transition,  $\nu$  and  $\beta$ . In order to do so, we use finite size scaling theory [77, 78]. Close the critical point,  $\langle \hat{m}_s^2 \rangle$  scales with the system size,  $N$ , as

$$\langle \hat{m}_s^2(N, J) \rangle = N^{-2\beta/\nu} f \left( N^{1/\nu} (J - J_c) \right), \quad (16)$$

where  $\nu$  is the critical exponent indicating how the correlation length diverges and  $\beta$  the exponent indicating how the magnetization diverges in the thermodynamic limit.  $f$  is a dimensionless scaling function that depends on the ratio between system size and correlation length and controls for finite size effects. The critical values are fitted so that different sizes collapse with the same  $f$ . In our case, we use the Python FSSA library [79, 80] for the fit. All the collapses obtained are plotted in Appendix A. Table II summarizes the critical values obtained for the most paradigmatic cases, comparing them to analytical results, where available, and to other numerical values reported in Refs. [43] and [47]. Note that, to enable a direct comparison, we do not apply the Kac renormalization factor of Eq. (4). In addition to  $J_c$ ,  $\nu$  and  $\beta$ , we also list  $\tilde{h}_c = 1/\tilde{J}_c$  and  $\theta_c = \arctan(1/\tilde{h}_c)$ .

FM						
$\alpha$	Method	$J_c$	$\tilde{h}_c$	$\theta_c$	$\nu$	$\beta$
1.5	Ours	-1.27(3)	4.57(9)	0.22(9)	1.6(5)	0.46(7)
	Ref. 43	-1.2107	4.7600	0.2071	1.9911	0.4968
	Theory	-	-	-	2.0	0.5
2.5	Ours	-2.09(1)	1.76(1)	0.516(9)	1.08(9)	0.19(3)
	Ref. 43	-2.0878	1.7631	0.5159	1.1084	0.1880
6.0	Ours	-2.963(1)	1.0242(4)	0.7734(2)	1.00(3)	0.12(3)
	Ref. 43	-2.9444	1.0307	0.7703	0.9849	0.1238
	Theory	-	-	-	1.0	0.125
AFM						
$\alpha$	Method	$J_c$	$\tilde{h}_c$	$\theta_c$	$\nu$	$\beta$
2.5	Ours	4.7527(6)	0.77449(9)	0.91180(3)	1.17(2)	0.0911(9)
	Ref. 43	4.6638	0.7893	0.9026	0.9103	0.1153
	Ref. 47	4.6778	0.7869	0.9041	1.01	0.122
6.0	Ours	3.143(2)	0.9655(5)	0.8029(2)	1.04(2)	0.11(2)
	Theory	-	-	-	1.0	0.125

TABLE II. List with all critical points and exponents obtained with a ViT ansatz through finite-size scaling analysis (marked as Ours in the Method column) for different values of  $\alpha$  and in both interaction regimes, FM ( $J_c < 0$ ) and AFM ( $J_c > 0$ ). Moreover, we show the analytical results when they exist, at  $\alpha \leq \alpha_{MF} = 5/3$  and  $\alpha = 6 > \alpha^*$ , as well as different values reported in Refs. 43 and 47.

The values presented in Table II are comparable to those found in the literature or indicated by the theory. During our investigation, we observed that the results

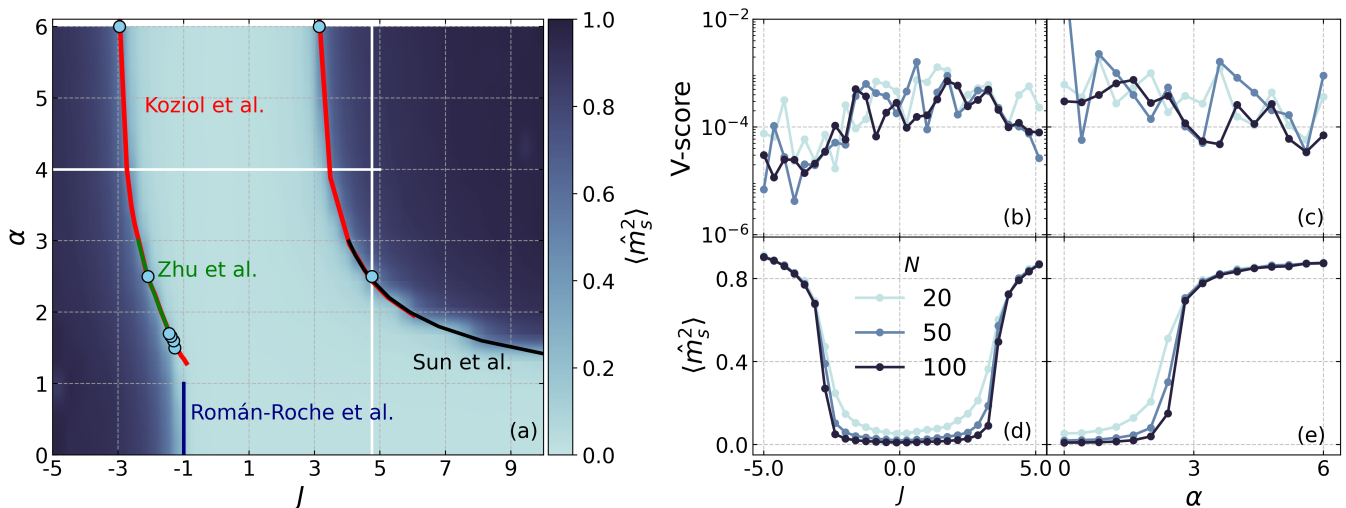


FIG. 4. (a) Squared magnetization,  $\langle \hat{m}_s^2 \rangle$ , obtained from the ground states determined by the ViT ansatz for a size of  $N = 50$  spins for different values of  $J$  and  $\alpha$ . In this diagram different transition curves present in the literature are represented [43–46]. In addition, the points indicated with light blue circles correspond to our estimates obtained by finite-size scaling analysis. The panels to the right of the phase diagram indicate the behavior of different quantities along two slices of the diagram (white lines). In panels (b) and (d) the value of  $\alpha = 4.0$  has been set and the V-score (15) and squared magnetization  $\langle \hat{m}_s^2 \rangle$  are shown as a function of  $J$ , respectively. In panels (c) and (e) the same quantities are shown in the same order but fixing  $J = 4.75$  and varying  $\alpha$ . The different curves in all these panels indicate different chain sizes, with  $N = 20$  being the lightest color and  $N = 100$  the darkest, going through  $N = 50$  (intermediate).

worsen for the smallest  $\alpha$ , which is unexpected since we are approaching the MF regime where the wave function should be easier to find. Therefore, we attribute this to finite-size effects. This was confirmed by simulating longer chains ( $N = 200$ ) and observing an improvement in the results.

Once the critical point has been determined, we can study the behavior of the entanglement entropy. On one hand, it reveals the limitations in the amount of entanglement other methods such as MPS can capture. On the other hand, it also characterizes the critical points, where it is known that the behavior of the entropy is universal and violates the area law, following a logarithmic dependence with the size  $N$  [81]. Our results for the entropy are plotted in Figure 5 as a function of  $J$  for both ferromagnetic and antiferromagnetic cases at  $\alpha = 2.5$ . These points correspond to single-shot executions. To aid visualization, we applied a Savitzky-Golay filter, which allows us to observe a certain trend across the different system sizes. Far from the critical point, indicated by a solid gray line in each case corresponding to the values reported in Table II, it is clear that the entropy becomes size-independent, in accordance with the area law. However, near the critical point, these data do not allow us to deduce the predicted volume dependence. Therefore, we focus on the estimated critical point and perform statistics over 10 independent runs. The insets in each of the panels of Figure 5 show the results for the mean and standard deviation. In this case, it is clear that the data are consistent with the expected logarithmic growth, validating the ansatz’s ability to reproduce highly entangled

states.

### C. Comparison to the RBM-like solution

We conclude this section by comparing the ViT and RBM-like architectures using the V-score (15) and limiting the total training time for each network to 3 minutes. The results for  $\alpha = 2.5$  and  $N = 50$  as a function of  $J$  are shown in Figure 6. The training used for the ViT is the same as the one described in Sec. III C. For the RBM, we use hand-optimized parameters for all values of  $J$ . However, since the chosen training protocol may be a crucial factor in the model’s performance, we explore two different protocols to ensure fairness in the comparison. These two protocols consist in different schedules for the learning rate  $\lambda$ : a linear decrease and a linear warm-up followed by an exponential decay. The first one corresponds to a linear decrease where the initial and final values of the learning rate are  $\lambda_0 = 0.1$  and  $\lambda_{\min} = 0.01$ , respectively. The other has the same structure as the one detailed in Sec. III C, with values for those parameters of  $\lambda_0 = 0.01$ ,  $\lambda_{\max} = 1.2$ ,  $n_{\text{warm}} = 50$  and  $\gamma = 0.995$ . The corresponding results for each protocol are shown in panels (a) and (b) of Figure 6, respectively. To allow for more iterations within the maximum allotted time, the number of MC samples has been reduced to 2048 in all cases.

The main observation from Figure 6 is that the ViT architecture achieves better quantum states across most of the spectrum in all scenarios, including within the critical



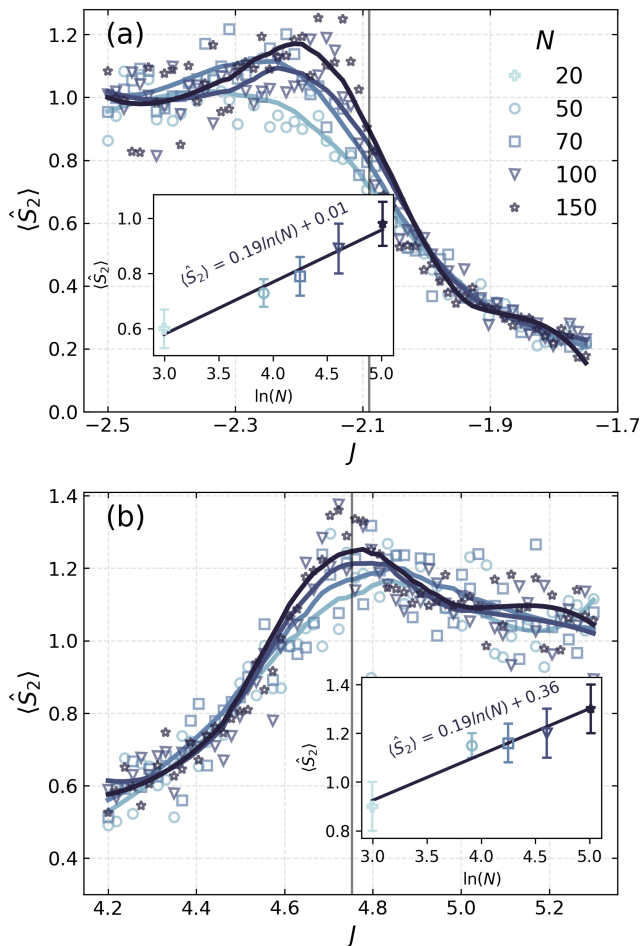


FIG. 5. Determination of the Renyi-2 entropy. For  $\alpha = 2.5$ , the entropy for different chain sizes is shown in the FM regime [panel (a)] and in the AFM regime [panel (b)]. The dots in the main panels represent the simulation results [cf. Eq. (13)] while the solid lines have been obtained with a Savitzky-Golay filter to act as a guide to the eye. The insets show the statistical analysis performed at the critical points of each case (vertical lines in the main panels). The mean value obtained from 10 independent runs is displayed, along with the corresponding standard deviation as error bars. The linear fit indicates that the system’s entropy follows the expected logarithmic scale in both cases.

regions, which are marked by dashed lines. In this regard, it is particularly interesting how the V-score worsens in the vicinity of the critical point for all the architectures and training protocols considered. This behaviour is similar to what has been observed in other ansätze used to obtain ground states such as quantum circuits [82]. Additionally, we see that increasing the RBM’s density briefly improves the results, but ultimately reaches a plateau in the V-score that cannot be surpassed by simply adding more hidden neurons to the RBM. The only region where all architectures converge in accuracy is around  $J = 0$ , where the model is trivial. In fact, numerical tests indicate that in this limit it is favourable to reduce the token

dimension to a single spin, replicating the non-interacting spin scenario, which in turn helps us understand why the simpler RBM achieves better accuracy here. Nonetheless, the ViT provides consistent quality results regardless of the value of  $J$  and the phase, whether ordered or not. Moreover, the RBM with  $\alpha_{\text{RBM}} = 1$  already has more parameters than our ViT architecture, with 5100 vs. 1133, respectively. Therefore, the ViT is able to achieve better results using fewer parameters and less training time, demonstrating its superiority in terms of efficiency.

Despite these results for the finite-size case, it is worth noting that a recent study [83] demonstrates that the RBM is capable of finding the ground state using only a single variational parameter in the thermodynamic limit ( $N \rightarrow \infty$ ) and in the strong long-range regime ( $\alpha < 1$ ).

## V. CONCLUSIONS

In this work, we explored the application of the Vision Transformer (ViT) to quantum long-range models, with a focus on the transverse-field Ising model. Our results affirm the ViT’s capability to handle all interaction regimes, from short-range to strong long-range interactions. Figure 4 summarizes the full phase diagram and its comparison with previous results that focus on specific regions. Additionally, Table II quantitatively compares the critical properties,  $J_c$ , and the critical exponents  $\nu$  and  $\beta$  with values reported in the literature.

Throughout this work, our intention has been to employ an economical approach in terms of the number of hyperparameters and to use single-shot results, without further curation or cherry-picking. The attention mechanism, a fundamental aspect of this architectural design, is responsible for identifying the most significant correlations within a relatively small number of iterations. Additionally, the remaining MLP layers endow the model with sufficient expressiveness and versatility, allowing it to produce a solution for any point within the parameter space. This is further corroborated by Figure 6, where a comparison was made with an ansatz based on RBMs setting the maximum training wall time. The ViT outperforms the accuracy obtained in critical regions by up to two orders of magnitude. Moreover, the use of the log-cosh function as a nonlinear activation function at the end of the core block [cf. Figure 3] was revealed as crucial: its inclusion, as described in Section III B, was fundamental in obtaining consistently and stably accurate states in the deep FM regime.

Our results illustrate the power of modern NN architectures to afford superior accuracy and efficiency than their more traditional counterparts when used as ansätze in variational quantum Monte Carlo workflows, and specifically demonstrate the ability of the ViT to capture important correlations in a cost-effective manner, opening the door to its use for a wider variety of problems involving long-range interactions.

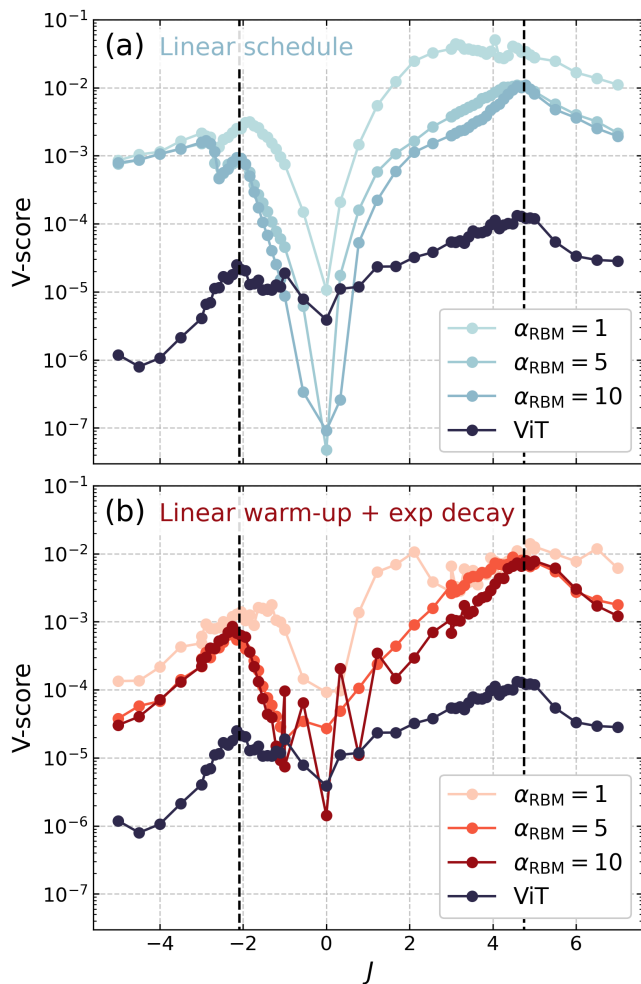


FIG. 6. Performance comparison between RBM-like models and ViT. Two different schedules for the learning rate have been used for the RBM, indicated in panels (a) and (b). The training for the ViT model is the same in both cases and it is described in Sec. III C. Fixing a maximum training time (3 minutes) for each point, the V-scores obtained by an RBM-like ansatz with different densities (denoted as  $\alpha_{\text{RBM}}$  in the legend) are compared with the ones obtained by our ViT implementation. The chain size is  $N = 50$  and the interaction range  $\alpha = 2.5$ . The black dashed lines mark the critical points obtained by FSSA [see Table II].

#### ACKNOWLEDGEMENTS

The authors acknowledge funding through grants CEX2023-001286-S, from

MCIN/AEI/10.13039/501100011033 and the EU “NextGenerationEU”/PRTR, and TED2021-131447B-C21, from MCIN/AEI/10.13039/501100011033. This study was also supported by MCIN with funding from European Union NextGenerationEU (PRTR-C17.I1) promoted by the Government of Aragon. We also acknowledge the Gobierno de Aragón (Grant E09-17R Q-MAD), Quantum Spain and the CSIC Quantum Technologies Platform PTI-001. This research project was made possible through the access granted by the Galician Supercomputing Center (CESGA) to its supercomputing infrastructure. The supercomputer FinisTerra III and its permanent data storage system have been funded by the Spanish Ministry of Science and Innovation, the Galician Government and the European Regional Development Fund (ERDF). S. R.-J. acknowledges financial support from Gobierno de Aragón through a doctoral fellowship.

#### Appendix A: Finite-size scaling analysis

Here we show all the collapses of the curves corresponding to the fluctuations that are used to perform finite-size scaling analysis (FSSA), as detailed in the main text, in order to obtain the critical parameters of the system. These curves are shown in Figure 7. For each value of  $\alpha$  considered, 60 points have been simulated in different windows for  $J$ . Not every point has been used to obtain the critical parameters; narrower windows centred on the critical point and different for each  $\alpha$  have been chosen in order to obtain the best collapses. However, in Figure 7 all the 60 points are shown. To get these points a maximum number of 500 iterations was set in the training procedure. The schedule for the SR parameter,  $\Delta_{\text{sr}}$ , the number of warm-up iterations,  $n_{\text{warm}}$ , as well as the decay rate  $\gamma$ , are all the same as the ones described in Section III C. The peak value for the learning rate,  $\lambda_{\text{max}}$  goes from 5.0 for the training of the smallest chains ( $N = 50$ ) up to 12.0 for the largest ( $N = 200$ ) taking intermediate values for the intermediate sizes.

The maximum chain size was  $N = 150$  spins for all interaction ranges except for  $\alpha = 1.5$ , where we made a further push to simulate  $N = 200$  spins to verify the finite size effects in the MF regime mentioned in the main text.

- 
- [1] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, *Ann. Phys.* **349**, 117–158 (2014).  
 [2] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, Dynamical mean-field theory of strongly correlated

- fermion systems and the limit of infinite dimensions, *Rev. Mod. Phys.* **68**, 13 (1996).  
 [3] F. Becca and S. Sorella, *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University Press, 2017).

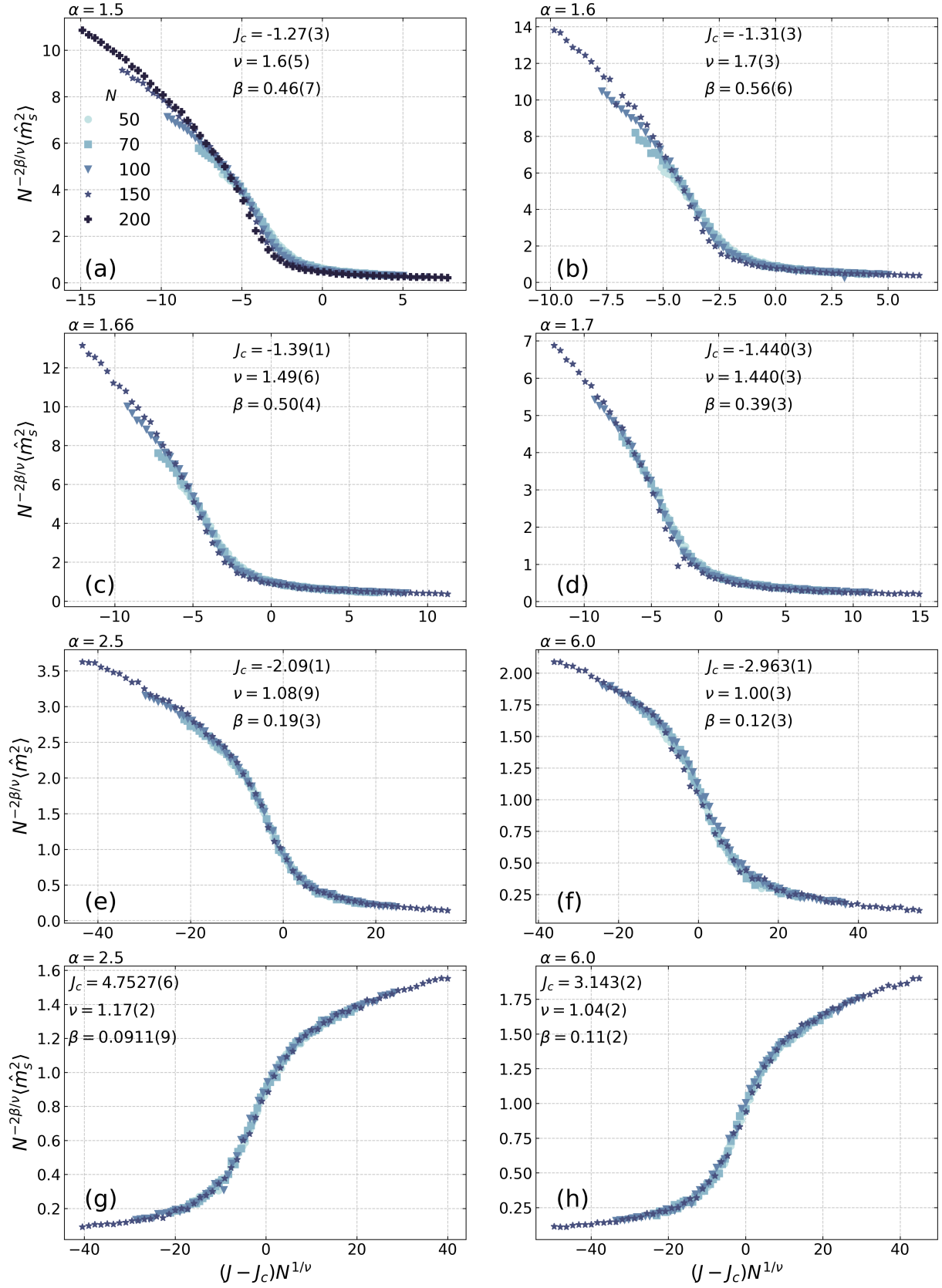


FIG. 7. Collapsed curves using the FSSA technique for the data obtained from the fluctuations  $\langle \hat{m}_s^2(N, J) \rangle$  for different values of  $\alpha$ . In each panel the values of the critical parameters obtained through the corresponding fit are shown.

- [4] Y. Alexeev, M. Amsler, P. Baity, M. A. Barroca, S. Bassini, T. Battelle, D. Camps, D. Casanova, Y. jai Choi, F. T. Chong, C. Chung, C. Codella, A. D. Corcoles, J. Cruise, A. D. Meglio, J. Dubois, I. Duran, T. Eckl, S. Economou, S. Eidenbenz, B. Elmegreen, C. Fare, I. Faro, C. S. Fernández, R. N. B. Ferreira, K. Fuji, B. Fuller, L. Gagliardi, G. Galli, J. R. Glick, I. Gobbi, P. Gokhale, S. de la Puente Gonzalez, J. Greiner, B. Gropp, M. Grossi, E. Gull, B. Healy, B. Huang, T. S. Humble, N. Ito, A. F. Izmaylov, A. Javadi-Abhari, D. Jennewein, S. Jha, L. Jiang, B. Jones, W. A. de Jong, P. Jurcevic, W. Kirby, S. Kister, M. Kitagawa, J. Klassen, K. Klymko, K. Koh, M. Kondo, D. M. Kurcuoglu, K. Kurowski, T. Laino, R. Landfield, M. Leininger, V. Leyton-Ortega, A. Li, M. Lin, J. Liu, N. Lorente, A. Luckow, S. Martiel, F. Martin-Fernandez, M. Martonosi, C. Marvinnay, A. C. Medina, D. Merten, A. Mezzacapo, K. Michielsen, A. Mitra, T. Mittal, K. Moon, J. Moore, M. Motta, Y.-H. Na, Y. Nam, P. Narang, Y. ya Ohnishi, D. Ottaviani, M. Otten, S. Pakin, V. R. Pascuzzi, E. Penault, T. Piontek, J. Pitera, P. Rall, G. S. Ravi, N. Robertson, M. Rossi, P. Rydlichowski, H. Ryu, G. Samsonidze, M. Sato, N. Saurabh, V. Sharma, K. Sharma, S. Shin, G. Slessman, M. Steiner, I. Suidikov, I.-S. Suh, E. Switzer, W. Tang, J. Thompson, S. Todo, M. Tran, D. Trenev, C. Trott, H.-H. Tseng, E. Tureci, D. G. Valinas, S. Vallecorsa, C. Wever, K. Wojciechowski, X. Wu, S. Yoo, N. Yoshioka, V. W. zhe Yu, S. Yunoki, S. Zhuk, and D. Zubarev, Quantum-centric supercomputing for materials science: A perspective on challenges and future directions (2023), [arXiv:2312.09733](https://arxiv.org/abs/2312.09733) [quant-ph].
- [5] H. J. Changlani, J. M. Kinder, C. J. Umrigar, and G. K.-L. Chan, Approximating strongly correlated wave functions with correlator product states, *Phys. Rev. B* **80**, 245116 (2009).
- [6] S. Anders, M. B. Plenio, W. Dür, F. Verstraete, and H.-J. Briegel, Ground-state approximation for strongly interacting spin systems in arbitrary spatial dimension, *Phys. Rev. Lett.* **97**, 107206 (2006).
- [7] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602–606 (2017).
- [8] D.-L. Deng, X. Li, and S. Das Sarma, Quantum entanglement in neural network states, *Phys. Rev. X* **7**, 021021 (2017).
- [9] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Neural-network quantum states, string-bond states, and chiral topological states, *Phys. Rev. X* **8**, 011006 (2018).
- [10] M. Medvidović and J. R. Moreno, [Neural-network quantum states for many-body physics](https://arxiv.org/abs/2402.11014) (2024), [arXiv:2402.11014](https://arxiv.org/abs/2402.11014) [cond-mat.dis-nn].
- [11] F. Cartarius, G. Morigi, and A. Minguzzi, Structural transitions of nearly second order in classical dipolar gases, *Phys. Rev. A* **90**, 053601 (2014).
- [12] D. S. L. Abergel and T. Chakraborty, Long-range coulomb interaction in bilayer graphene, *Phys. Rev. Lett.* **102**, 056807 (2009).
- [13] L. Béguin, A. Vernier, R. Chicireanu, T. Lahaye, and A. Browaeys, Direct measurement of the van der waals interaction between two rydberg atoms, *Phys. Rev. Lett.* **110**, 263201 (2013).
- [14] J. W. Britton, B. C. Sawyer, A. C. Keith, C.-C. J. Wang, J. K. Freericks, H. Uys, M. J. Biercuk, and J. J. Bollinger, Engineered two-dimensional ising interactions in a trapped-ion quantum simulator with hundreds of spins, *Nature* **484**, 489 (2012).
- [15] M. Knap, A. Kantian, T. Giamarchi, I. Bloch, M. D. Lukin, and E. Demler, Probing real-space and time-resolved correlation functions with many-body ramsey interferometry, *Phys. Rev. Lett.* **111**, 147205 (2013).
- [16] C. Monroe, W. C. Campbell, L. M. Duan, Z. X. Gong, A. V. Gorshkov, P. W. Hess, R. Islam, K. Kim, N. M. Linke, G. Pagano, P. Richerme, C. Senko, and N. Y. Yao, Programmable quantum simulations of spin systems with trapped ions, *Rev. Mod. Phys.* **93**, 025001 (2021).
- [17] A. Browaeys and T. Lahaye, Many-body physics with individually controlled rydberg atoms, *Nat. Phys.* **16**, 132 (2020).
- [18] D. Mukamel, Statistical mechanics of systems with long range interactions, *AIP Conf. Proc.* **970**, 22 (2008).
- [19] A. Campa, T. Dauxois, and S. Ruffo, Statistical mechanics and dynamics of solvable models with long-range interactions, *Phys. Rep.* **480**, 57 (2009).
- [20] S. Fey and K. P. Schmidt, Critical behavior of quantum magnets with long-range interactions in the thermodynamic limit, *Phys. Rev. B* **94**, 075156 (2016).
- [21] N. Defenu, T. Donner, T. Macrì, G. Pagano, S. Ruffo, and A. Trombettoni, Long-range interacting quantum systems, *Rev. Mod. Phys.* **95**, 035002 (2023).
- [22] N. Defenu, A. Leroze, and S. Pappalardi, Out-of-equilibrium dynamics of quantum many-body systems with long-range interactions, *Phys. Rep.* **1074**, 1 (2024), out-of-equilibrium dynamics of quantum many-body systems with long-range interactions.
- [23] T. Koffel, M. Lewenstein, and L. Tagliacozzo, Entanglement entropy for the long-range ising chain in a transverse field, *Phys. Rev. Lett.* **109**, 267203 (2012).
- [24] T. Kuwahara and K. Saito, Area law of noncritical ground states in 1d long-range interacting systems, *Nat. Commun.* **11**, 4478 (2020).
- [25] D. Vodola, L. Lepori, E. Ercolessi, A. V. Gorshkov, and G. Pupillo, Kitaev chains with long-range pairing, *Phys. Rev. Lett.* **113**, 156402 (2014).
- [26] F. Ares, J. G. Esteve, F. Falceto, and A. R. de Queiroz, Entanglement entropy in the long-range kitaev chain, *Phys. Rev. A* **97**, 062301 (2018).
- [27] D. Vodola, L. Lepori, E. Ercolessi, and G. Pupillo, Long-range ising and kitaev models: phases, correlations and edge modes, *New J. Phys.* **18**, 015001 (2015).
- [28] L. Vanderstraeten, M. VanDamme, H. P. Büchler, and F. Verstraete, Quasiparticles in quantum spin chains with long-range interactions, *Phys. Rev. Lett.* **121**, 090603 (2018).
- [29] G. Francica and L. Dell'Anna, Correlations, long-range entanglement, and dynamics in long-range kitaev chains, *Phys. Rev. B* **106**, 155126 (2022).
- [30] J. T. Schneider, J. Despres, S. J. Thomson, L. Tagliacozzo, and L. Sanchez-Palencia, Spreading of correlations and entanglement in the long-range transverse ising chain, *Phys. Rev. Res.* **3**, L012022 (2021).
- [31] S. B. Jäger, L. Dell'Anna, and G. Morigi, Edge states of the long-range kitaev chain: An analytical study, *Phys. Rev. B* **102**, 035152 (2020).
- [32] O. Viyuela, D. Vodola, G. Pupillo, and M. A. Martin-Delgado, Topological massive dirac edge modes and long-range superconducting hamiltonians, *Phys. Rev. B* **94**,

- 125121 (2016).
- [33] H. Saito, Solving the bose–hubbard model with machine learning, *J. Phys. Soc. Jpn.* **86**, 093001 (2017).
- [34] 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).
- [35] 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).
- [36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017).
- [37] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research* **21**, 1 (2020).
- [38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale (2021), [arXiv:2010.11929 \[cs.CV\]](https://arxiv.org/abs/2010.11929).
- [39] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, Highly accurate protein structure prediction with alphafold, *Nature* **596**, 583 (2021).
- [40] L. L. Viteritti, R. Rende, and F. Becca, Transformer variational wave functions for frustrated quantum spin systems, *Phys. Rev. Lett.* **130**, 236401 (2023).
- [41] K. Sprague and S. Czischek, Variational monte carlo with large patched transformers, *Commun. Phys.* **7**, 90 (2024).
- [42] D. Luo, Z. Chen, K. Hu, Z. Zhao, V. M. Hur, and B. K. Clark, Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models, *Phys. Rev. Res.* **5**, 013216 (2023).
- [43] J. A. Koziol, A. Langheld, S. C. Kapfer, and K. P. Schmidt, Quantum-critical properties of the long-range transverse-field ising model from quantum monte carlo simulations, *Phys. Rev. B* **103**, 245135 (2021).
- [44] Z. Zhu, G. Sun, W.-L. You, and D.-N. Shi, Fidelity and criticality of a quantum ising chain with long-range interactions, *Phys. Rev. A* **98**, 023607 (2018).
- [45] J. Román-Roche, V. Herráiz-López, and D. Zueco, Exact solution for quantum strong long-range models via a generalized hubbard-stratonovich transformation, *Phys. Rev. B* **108**, 165130 (2023).
- [46] G. Sun, Fidelity susceptibility study of quantum long-range antiferromagnetic ising chain, *Phys. Rev. A* **96**, 043621 (2017).
- [47] J. Kim, D. Kim, and D.-H. Kim, Neural-network quantum-state study of the long-range antiferromagnetic ising chain, *Phys. Rev. E* **109**, 064123 (2024).
- [48] A. Campa, T. Dauxois, D. Fanelli, and S. Ruffo, *Physics of Long-Range Interacting Systems* (Oxford University Press, 2014).
- [49] E. G. Lazo, M. Heyl, M. Dalmonte, and A. Angelone, Finite-temperature critical behavior of long-range quantum ising models, *SciPost Phys.* **11**, 076 (2021).
- [50] A. Campa, A. Giansanti, and D. Moroni, Canonical solution of a system of long-range interacting rotators on a lattice, *Phys. Rev. E* **62**, 303 (2000).
- [51] T. Mori, Instability of the mean-field states and generalization of phase separation in long-range interacting systems, *Phys. Rev. E* **84**, 031128 (2011).
- [52] T. Mori, Analysis of the exactness of mean-field theory in long-range interacting systems, *Phys. Rev. E* **82**, 060103(R) (2010).
- [53] T. Mori, Microcanonical analysis of exactness of the mean-field theory in long-range interacting systems, *J. Stat. Phys.* **147**, 1020 (2012).
- [54] T. Mori, Equilibrium properties of quantum spin systems with nonadditive long-range interactions, *Phys. Rev. E* **86**, 021132 (2012).
- [55] A. Dutta and J. K. Bhattacharjee, Phase transitions in the quantum ising and rotor models with a long-range interaction, *Phys. Rev. B* **64**, 184106 (2001).
- [56] N. Defenu, A. Trombettoni, and S. Ruffo, Criticality and phase diagram of quantum long-range  $o(n)$  models, *Phys. Rev. B* **96**, 104432 (2017).
- [57] I. Kolokolov, Functional integration for quantum magnets: New method and new results, *Ann. Phys.* **202**, 165 (1990).
- [58] P. Schauß, J. Zeiher, T. Fukuhara, S. Hild, M. Cheneau, T. Macrì, T. Pohl, I. Bloch, and C. Gross, Crystallization in ising quantum magnets, *Science* **347**, 1455–1458 (2015).
- [59] H. Labuhn, D. Barredo, S. Ravets, S. de Léséleuc, T. Macrì, T. Lahaye, and A. Browaeys, Tunable two-dimensional arrays of single rydberg atoms for realizing quantum ising models, *Nature* **534**, 667–670 (2016).
- [60] S. Sorella and L. Capriotti, Green function monte carlo with stochastic reconfiguration: An effective remedy for the sign problem, *Phys. Rev. B* **61**, 2599 (2000).
- [61] F. Vicentini, D. Hofmann, A. Szabó, D. Wu, C. Roth, C. Giuliani, G. Pescia, J. Nys, V. Vargas-Calderón, N. Astrakhantsev, and G. Carleo, NetKet 3: Machine Learning Toolbox for Many-Body Quantum Systems, *SciPost Phys. Codebases*, 7 (2022).
- [62] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signal.* **2**, 303 (1989).
- [63] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks* **2**, 359 (1989).
- [64] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, USA, 2016).
- [65] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift* (2015), [arXiv:1502.03167 \[cs.LG\]](https://arxiv.org/abs/1502.03167).
- [66] J. L. Ba, J. R. Kiros, and G. E. Hinton, Layer normalization, [arXiv:1607.06450 \[stat.ML\]](https://arxiv.org/abs/1607.06450), [arXiv:1607.06450 \[stat.ML\]](https://arxiv.org/abs/1607.06450).
- [67] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) pp. 770–778.
- [68] D. Chen, F. Hu, G. Nian, and T. Yang, Deep residual learning for nonlinear regression, *Entropy* **22**,

- 10.3390/e22020193 (2020).
- [69] P. Ramachandran, B. Zoph, and Q. V. Le, Searching for activation functions (2017), [arXiv:1710.05941](https://arxiv.org/abs/1710.05941) [cs.NE].
- [70] P. Shaw, J. Uszkoreit, and A. Vaswani, Self-attention with relative position representations, [CoRR abs/1803.02155](https://arxiv.org/abs/1803.02155) (2018), 1803.02155.
- [71] [https://github.com/Theory-and-Simulation-at-INMA/transformer\\_LR\\_WF\\_public/tree/v24.07.1](https://github.com/Theory-and-Simulation-at-INMA/transformer_LR_WF_public/tree/v24.07.1), doi:10.5281/zenodo.12654460, retrieved on 2024-07-04.
- [72] J. Heek, A. Levskaya, A. Oliver, M. Ritter, B. Rondepierre, A. Steiner, and M. van Zee, Flax: A neural network library and ecosystem for JAX (2020), <http://github.com/google/flax>, accessed on 2021-12-09.
- [73] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: composable transformations of Python+NumPy programs (2018), <http://github.com/google/jax>, accessed on 2021-12-09.
- [74] DeepMind, I. Babuschkin, K. Baumli, A. Bell, S. Bhupatiraju, J. Bruce, P. Buchlovsky, D. Budden, T. Cai, A. Clark, I. Danihelka, A. Dedieu, C. Fantacci, J. Godwin, C. Jones, R. Hemsley, T. Hennigan, M. Hessel, S. Hou, S. Kapturovski, T. Keck, I. Kemaev, M. King, M. Kunesch, L. Martens, H. Merzic, V. Mikulik, T. Norman, G. Papamakarios, J. Quan, R. Ring, F. Ruiz, A. Sanchez, L. Sartran, R. Schneider, E. Sezener, S. Spencer, S. Srinivasan, M. Stanojević, W. Stokowiec, L. Wang, G. Zhou, and F. Viola, [The DeepMind JAX Ecosystem](https://arxiv.org/abs/2006.10271) (2020).
- [75] M. B. Hastings, I. González, A. B. Kallin, and R. G. Melko, Measuring renyi entanglement entropy in quantum monte carlo simulations, [Phys. Rev. Lett.](https://arxiv.org/abs/1005.3548) **104**, 157201 (2010).
- [76] D. Wu, R. Rossi, F. Vicentini, N. Astrakhantsev, F. Becca, X. Cao, J. Carrasquilla, F. Ferrari, A. Georges, M. Hibat-Allah, M. Imada, A. M. Läuchli, G. Mazzola, A. Mezzacapo, A. Millis, J. R. Moreno, T. Neupert, Y. Nomura, J. Nys, O. Parcollet, R. Pohle, I. Romero, M. Schmid, J. M. Silvester, S. Sorella, L. F. Tocchio, L. Wang, S. R. White, A. Wietek, Q. Yang, Y. Yang, S. Zhang, and G. Carleo, Variational benchmarks for quantum many-body problems (2023), [arXiv:2302.04919](https://arxiv.org/abs/2302.04919) [quant-ph].
- [77] M. E. Newman and G. T. Barkema, *Monte Carlo methods in statistical physics* (Clarendon Press, 1999).
- [78] K. Binder, D. M. Ceperley, J.-P. Hansen, M. Kalos, D. Landau, D. Levesque, H. Mueller-Krumbhaar, D. Stauffer, and J.-J. Weis, *Monte Carlo methods in statistical physics*, Vol. 7 (Springer Science & Business Media, 2012).
- [79] A. Sorge, [pyfssa 0.7.6](https://pypi.org/project/pyfssa/) (2015).
- [80] O. Melchert, autoscale.py - a program for automatic finite-size scaling analyses: A user's guide (2009), [arXiv:0910.5403](https://arxiv.org/abs/0910.5403) [physics.comp-ph].
- [81] P. Calabrese and J. Cardy, Entanglement entropy and conformal field theory, [Journal of Physics A: Mathematical and Theoretical](https://arxiv.org/abs/hep-th/0411177) **42**, 504005 (2009).
- [82] S. Roca-Jerat, T. Sancho-Lorente, J. Román-Roche, and D. Zueco, Circuit Complexity through phase transitions: Consequences in quantum state preparation, [SciPost Phys.](https://arxiv.org/abs/2208.02131) **15**, 186 (2023).
- [83] F. B. Trigueros, T. Mendes-Santos, and M. Heyl, Simplicity of mean-field theories in neural quantum states, [Phys. Rev. Res.](https://arxiv.org/abs/2401.02326) **6**, 023261 (2024).