Solving Quantum Many-Particle Models with Graph Attention Network

Qi-Hang Yu(于启航)[1] and Zi-Jing Lin(林子敬)[1,2*]

[1]*Department of Physics, University of Science and Technology of China, Hefei 230026, China*
[2]*Hefei National Laboratory, University of Science and Technology of China, Hefei 230088, China*

(Received 7 January 2024; accepted manuscript online 4 March 2024)

Deep learning methods have been shown to be effective in representing ground-state wavefunctions of quantum many-body systems, however the existing approaches cannot be easily used for non-square like or large systems. Here, we propose a variational ansatz based on the graph attention network (GAT) which learns distributed latent representations and can be used on non-square lattices. The GAT-based ansatz has a computational complexity that grows linearly with the system size and can be extended to large systems naturally. Numerical results show that our method achieves the state-of-the-art results on spin-1/2 J_1–J_2 Heisenberg models over the square, honeycomb, triangular, and kagome lattices with different interaction strengths and lattice sizes (up to 24×24 for square lattice). The method also provides excellent results for the ground states of transverse field Ising models on square lattices. The GAT-based techniques are efficient and versatile and hold promise for studying large quantum many-body systems with exponentially sized objects.

DOI: 10.1088/0256-307X/41/3/030202

The study of quantum many-body systems is essential in physics, but remains to be very difficult because of the exponential growth of Hilbert space dimensions with the number of particles in the systems. Over the last decades, the advent of sophisticated computational methods has led to breakthroughs in addressing quantum many-body problems. Tensor network[1–4] is one of the popular techniques to model quantum many-body systems but suffers entanglement problems.[5] Quantum Monte Carlo (QMC)[6–8] fails on many systems, mostly owing to the sign problem in it.[9] Variational Monte Carlo (VMC)[10,11] is more general for obtaining quantum many-body wavefunctions by optimizing a compact parameterized variational ansatz with self-sampled data. The central challenge of the VMC approach lies in designing a suitable variational ansatz that can accurately capture the structure of the lowest energy eigenstate.[11] Because it is extremely challenging to construct high-quality analytical approximations and can introduce strong biases, developing flexible parameterizations for a variety of problems is an interesting and important issue.

According to the universal approximation theorem,[12–14] in principle, a neural network can fit any function given sufficiently many parameters. Because of having favorable expressive power, neural network-based machine learning methods have been applied to solve various physical problems.[15–18] Recently, neural networks, such as restricted Boltzmann machines (RBMs), have been used as variational ansatz,[19] referred as neural quantum states (NQS). Following this direction, some studies explore deep Boltzmann machines[20] and fully connected neural networks to represent quantum states.[21] Many recent studies use convolutional neural networks (CNNs) as variational ansatz for square lattice systems[22–24] due to their image-like structures. On account of the advantages of autoregressive sampling, recurrent neural networks (RNNs)[25] are also used as variational ansatz.

Solving quantum frustrated models is challenging for numerical approaches, and has become an important way for evaluating new calculation methods. Some authors tackled the highly frustrated spin-1/2 J_1–J_2 model on the square lattice via CNNs.[22–24] They obtained ground state energies that were lower than the string-bond-states method[26] and plain Gutzwiller projected fermionic wavefunctions (GWF).[27] Ferrari *et al.*[28] studied the J_1–J_2 model by combining GWF with a complex-valued RBM correlator, they even got better results than those obtained by the density-matrix renormalization group (DMRG) method.[29] The best ground state energies in literature for $J_2/J_1 = 0.5$ on 10×10 square lattice are obtained by RBM[30] and transformer-based[31] NQS, which are lower than the previous best results obtained by physical inspired variational wavefunctions.[27] Compared to physical inspired wavefunctions and other numerical methods, these NQS greatly improve the accuracy and universality of VMC, and thus have great potential to solve a large class of quantum many-particle problems.

However, the above-mentioned NQS methods limit their applications to quantum many-body systems due to the following drawbacks. As a larger lattice size leads to smaller boundary effects, it is necessary for a simulation to capture the correct physics of quantum systems with long range correlations. Therefore, the computational scalability should be a criterion for selecting NQS methods. To consider an architecture scalable, both the number of its parameters and the network evaluation cost should at most grow mildly with the system size N.[32] The RBMs and its variants violate the first requirement because of a given constant ratio α between the number of hidden and visible neurons, the number of parameters grows quadratically.[30] RNNs violate the second requirement with its quadratic scaling computing cost, as the cost of

*Corresponding author. Email: zjlin@ustc.edu.cn
© 2024 Chinese Physical Society and IOP Publishing Ltd

evaluating an unsymmetrized configuration grows linearly with the system size N, and the symmetrization gives another factor N.[32] Meanwhile, the autoregressive property imposes specific constraints on the architecture of RNNs, thereby restricting their expressiveness.[33] Recently proposed transformer-based wavefunctions also come at a significant computational cost due to their quadratic complexity with respect to the input sequence length.[34] Finally, CNNs cannot be naturally applied to non-square like systems.[35–37]

In this Letter, we present an efficient and flexible variational model based on the graph attention network (GAT)[35] that can be applied to non-square lattices. The model learns distributed latent representations of complex wavefunction amplitudes that respect the underlying physical symmetries of the problem using an encode-process-decode architecture. Employing the attention mechanism allows us to better capture correlations between spins, which enhances the expressivity of the model. For common lattices, the number of trainable parameters does not change with the system size, and the computational complexity varies with the system size as $O(N)$. Thus, the model meets the requirements for scalability and allows us to use transfer learning to accelerate convergence. Compared to frequently used CNN, RBM and RNN-based methods, the GAT-based model can solve a variety of challenging problems on different lattices and large systems naturally. Numerical results show that our model achieves performance on par or better than state-of-the-art methods for the spin-1/2 J_1–J_2 Heisenberg models with various interaction strengths and lattices including square, honeycomb, triangular, and kagome lattices.

Variational Monte Carlo. A sketch of the VMC process is shown in Fig. 1. For a system of N spin-1/2 particles, each $\hat{\sigma}^z$ basis state $|\boldsymbol{\sigma}\rangle = |\sigma_1^z, \sigma_2^z, \ldots, \sigma_N^z\rangle$ is represented by N values in $\{\pm 1\}$ indicating the configuration of each spin. Given a variational model (commonly referred to as ansatz) $|\psi_\theta\rangle = \sum_{\boldsymbol{\sigma}} \psi_\theta(\boldsymbol{\sigma})|\boldsymbol{\sigma}\rangle$ and its parameters θ as shown in Fig. 1, the expectation value of Hamiltonian $\hat{H}$ can be written as

$$E(\theta) = \frac{\langle\psi_\theta|\hat{H}|\psi_\theta\rangle}{\langle\psi_\theta|\psi_\theta\rangle} = \sum_{\boldsymbol{\sigma}} p_\theta(\boldsymbol{\sigma}) E_{\mathrm{loc}}^\theta(\boldsymbol{\sigma}) = \langle\!\langle E_{\mathrm{loc}}^\theta \rangle\!\rangle_\theta, \quad (1)$$

where $p_\theta(\boldsymbol{\sigma}) \equiv |\psi_\theta(\boldsymbol{\sigma})|^2/\langle\psi_\theta|\psi_\theta\rangle$ can be seen as a probability distribution on spin configurations, $E_{\mathrm{loc}}^\theta(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma'}}\langle\boldsymbol{\sigma}|\hat{H}|\boldsymbol{\sigma'}\rangle\psi_\theta(\boldsymbol{\sigma'})/\psi_\theta(\boldsymbol{\sigma})$ is known as local energy, and $\langle\!\langle\cdot\rangle\!\rangle_\theta$ represents an expectation value with respect to $p_\theta(\boldsymbol{\sigma})$. The sum over all basis states is infeasible to evaluate for large N, and the Monte Carlo method is used to estimate the expectation efficiently. For the sampler module of Fig. 1, the samples satisfying the probability distribution $p_\theta(\boldsymbol{\sigma})$ are obtained by the Metropolis–Hastings algorithm, and the expectation value can be approximated by the empirical mean

$$\langle\!\langle E_{\mathrm{loc}}^\theta \rangle\!\rangle_\theta \simeq \frac{1}{N_{\mathrm{MC}}} \sum_{n=1}^{N_{\mathrm{MC}}} E_{\mathrm{loc}}^\theta(\boldsymbol{\sigma}^n), \quad (2)$$

where N_{MC} is the number of samples, and $\boldsymbol{\sigma}^n$ is the nth sample which satisfies the probability distribution $p_\theta(\boldsymbol{\sigma})$. The expectation value of any quantum operator can also be calculated like energy.

Then, in the updater of Fig. 1, logarithmic derivatives of variational wavefunction $\partial_\theta \log\psi_\theta(\boldsymbol{\sigma})$ and gradients of Hamiltonian expectation $\partial_\theta E(\theta)$ with respect to variational parameters θ are used to optimize the ansatz by either standard stochastic gradient methods or more advanced methods like the stochastic reconfiguration (SR). SR is a generic method for optimizing parameterized trial wavefunctions so as to minimize their energy with respect to a Hamiltonian $\hat{H}$.[11] This protocol has been shown to outperform stochastic gradient descent and other commonly used deep learning optimization algorithms in terms of stability and convergence for NQS.[38]

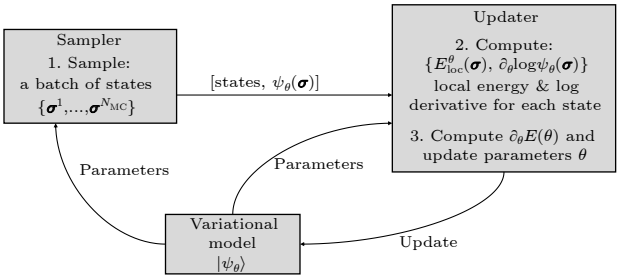

Fig. 1. The flow chart of VMC algorithm. Given a variational model $|\psi_\theta\rangle$, a batch of states satisfying $|\psi_\theta(\boldsymbol{\sigma})|^2/\langle\psi_\theta|\psi_\theta\rangle$ is sampled by Monte Carlo method in the sampler, and then used to calculate energy $E(\theta)$. Subsequently, logarithm derivatives $\partial_\theta \log\psi_\theta(\boldsymbol{\sigma})$ and gradients of energy $\partial_\theta E(\theta)$ are calculated to update parameters θ of $|\psi_\theta\rangle$ in the updater. The above process is repeated until convergence.

VMC is variational as it yields an upper bound for the energy of a system. Therefore, if two variational solutions give different energies, the one with the lower energy is closer to the true solution, even if the true energy and solution are not known.

Network Structure. We use a deep neural network to specify a variational wavefunction $\psi_\theta(\boldsymbol{\sigma})$. The network defines a complex-valued vector in the Hilbert space $|\psi_\theta\rangle = \sum_{\boldsymbol{\sigma}} \psi_\theta(\boldsymbol{\sigma})|\boldsymbol{\sigma}\rangle$, by establishing a mapping between the basis configuration $\boldsymbol{\sigma}$ and its corresponding complex amplitude $\psi_\theta(\boldsymbol{\sigma}) \in \mathbb{C}$. For convenience and numerical stability considerations,[39,40] we take the network to represent the logarithm of the wavefunction $\log\psi_\theta(\boldsymbol{\sigma})$ by predicting two scalar values: log modulus $\log(|\psi_\theta(\boldsymbol{\sigma})|)$ and phase $\arg(\psi_\theta(\boldsymbol{\sigma}))$ of its complex amplitude.

Our model uses GAT to compute $\log\psi_\theta(\boldsymbol{\sigma})$ from an input basis configuration $\boldsymbol{\sigma}$ in three steps as shown in Fig. 2. GAT was introduced to deal with arbitrary structured graphs,[35] and our GAT-based model can solve quantum many-body systems on non-square lattices, which cannot be handled by CNNs. First, an initial graph $G = (V, E)$ is constructed based on the input configuration $\boldsymbol{\sigma}$ and the geometry of the lattice, where $V = \{v_i\}$ and $E = \{\epsilon_{ij}\}$ are the sets of nodes and edges, respectively. To enhance the generalization ability of the GAT model and offer additional geometric and symmetric insights into the lattice, we incorporate sublattice encoding $\boldsymbol{s} = (s_1, \ldots, s_N)$ as an auxiliary input, following the approach described in Ref. [40]. The sublattice encoding s_i represents the position of node i in a unit cell that satisfies all the symmetries of the lattice, which has been successfully used in constructing variational states for classical[41] and quantum[42] systems. Examples of sublattice encodings used in this work

are illustrated in Fig. 3. Then nodes $v_i \in V$ are initialized by concatenating σ_i^z and s_i. Edge $\epsilon_{ij} \in E$ indicates that there is an edge between node i and node j, i.e., they are adjacent. To better capture interactions between the spin i and its next nearest neighbors, we assume the existence of edges connecting node i with both its nearest and next nearest neighbors.

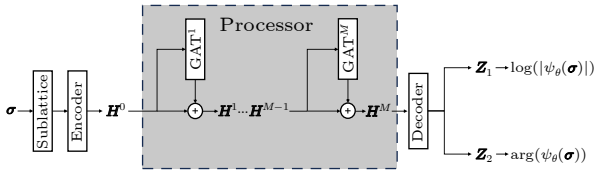

Fig. 2. Architecture of the GAT-based ansatz. Our model evaluates scalar outputs $\log(|\psi_\theta(\boldsymbol{\sigma})|)$ and $\arg(\psi_\theta(\boldsymbol{\sigma}))$ from an input basis element $\boldsymbol{\sigma}$ represented as a graph.

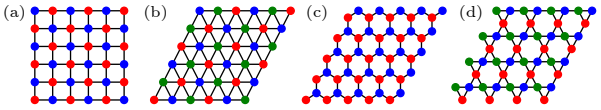

Fig. 3. Examples of sublattice encodings $\boldsymbol{s}$ used in this work as auxiliary input.

Afterwards, we utilize a GAT model with an encode-process-decode architecture[43] to compute a pair of final latent representations $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, which represent the real and imaginary parts, respectively. Instead of relying solely on a single latent representation $\boldsymbol{Z}$ to produce the wavefunction,[40] the utilization of two latent representations here effectively mitigates the coupling between the highly oscillatory phase and slow varying magnitude of the wavefunction to a certain extent.

The encoder maps nodes V of input graph G to a latent representation $\boldsymbol{H}^0 = \{\boldsymbol{h}_1^0, \boldsymbol{h}_2^0, \ldots, \boldsymbol{h}_N^0\}$ of width F by a multilayer perceptron (MLP).[44] These node features are updated by the processor that contains M rounds of message passing by GAT then. The input for each GAT layer is a set of node features $\boldsymbol{H}^t = \{\boldsymbol{h}_1^t, \boldsymbol{h}_2^t, \ldots, \boldsymbol{h}_N^t\}$, where $\boldsymbol{h}_i^t \in \mathbb{R}^F$ is the feature vector for spin i after t updates. This representation is updated through a message passing process in a residual manner[45] as shown in Fig. 4(a),

$$\boldsymbol{h}_i^{t+1} = \boldsymbol{h}_i^t + \mathrm{GAT}^{t+1}(\boldsymbol{h}_i^t, \{\boldsymbol{h}_j^t, \epsilon_{ij}\}), \tag{3}$$

where GAT^{t+1} is an update function that depends on the previous features $\boldsymbol{h}_i^t$ and neighboring node features $\{\boldsymbol{h}_j^t, \epsilon_{ij}\}$, ϵ_{ij} indicates that nodes i and j are adjacent. GAT^{t+1} outputs an intermediate representation $\tilde{\boldsymbol{H}}^{t+1} = \{\tilde{\boldsymbol{h}}_1^{t+1}, \tilde{\boldsymbol{h}}_2^{t+1}, \ldots, \tilde{\boldsymbol{h}}_N^{t+1}\}$, $\tilde{\boldsymbol{h}}_i^{t+1} \in \mathbb{R}^F$, that is used in residual connection. The complete GAT layer generates a new set of node features $\boldsymbol{H}^{t+1} = \{\boldsymbol{h}_1^{t+1}, \boldsymbol{h}_2^{t+1}, \ldots, \boldsymbol{h}_N^{t+1}\}$, $\boldsymbol{h}_i^{t+1} \in \mathbb{R}^F$.

Unlike other graph neural network (GNN)-based models,[36,40] our GAT layers employ a weighted attention mechanism which is used to inform models about the relative importance of different features for their respective tasks for the update functions. The attention mechanism is crucial for various cutting-edge sequence-to-sequence models employed in machine translation and language processing,[46,47] and has recently shown excellent results on graphs[35,48] and in some material sciences.[48,49] In recent quantum physics applications,[50,51] the attention mechanism allows the capture of complex quantum mechanical correlations between electrons or spins to achieve better results. In this work, we use the following self-attention mechanism to introduce spin-spin dependence into the variational wavefunction, thereby improving the expressive power of our model.

In order to achieve a higher-level representation, every GAT layer starts off with a shared node-wise linear transformation, parameterized by a weight matrix, $\boldsymbol{W}^{t+1} \in \mathbb{R}^{F\times F}$, the computational complexity of the transformation is $O(|V| \times F \times F)$, where $|V|$ is the number of nodes in the graph and increases as $O(N)$. After the transformation, the vectors $\boldsymbol{W}^{t+1}\boldsymbol{h}_i^t$ are recombined in some way at each node by performing self-attention mechanism on the nodes. This process dynamically computes attention coefficients e_{ij}^{t+1} across pairs of spins,

$$e_{ij}^{t+1} = f^{t+1}(\boldsymbol{W}^{t+1}\boldsymbol{h}_i^t || \boldsymbol{W}^{t+1}\boldsymbol{h}_j^t), \tag{4}$$

where f^{t+1} is implemented as a single-layer MLP, parameterized by a weight vector $\boldsymbol{a}^{t+1} \in \mathbb{R}^{2F}$, and applying the leaky rectified linear function (LeakyReLU) nonlinearity (with a slope $\alpha = 0.2$ for negative inputs) to add dependency on $\boldsymbol{h}_i^t$ to the attention. This weighted summation mechanism avoids the over-smoothing problem in GNNs.[52] Here f^{t+1} maps $2F$-dimensional vectors to real numbers, the attention coefficients are calculated for all edges in the graph, resulting in a computational complexity of $O(|E| \times F)$, where $|E|$ is the number of edges in the graph and increases as $O(N)$ for common lattices. To make the above coefficients easily comparable across different nodes, they are normalized across all choices of j using the softmax function

$$\alpha_{ij}^{t+1} = \mathrm{softmax}_j(e_{ij}^{t+1}) = \frac{\exp(e_{ij}^{t+1})}{\sum_{k\in\mathcal{N}_i}\exp(e_{ik}^{t+1})}. \tag{5}$$

We employ a masked attention mechanism, wherein we exclusively compute e_{ij}^{t+1} for nodes $j \in \mathcal{N}_i$ that are connected to node i through edges ϵ_{ij}. As mentioned above, to capture interactions between parts of second nearest neighbors and inject the lattice structure into GAT layers, we consider $\mathcal{N}_i$ as encompassing both first and second-order neighbors of node i, including node i itself.

Before combining features using the attention coefficients α_{ij}^{t+1}, a multi-head attention mechanism[46] as depicted in Fig. 4(b) is employed to stabilize the learning process of self-attention and improve the performance of GAT. To apply residual connection, we split the vectors $\boldsymbol{W}^{t+1}\boldsymbol{h}_i^t$ into K sub-arrays $(\boldsymbol{W}^{t+1}\boldsymbol{h}_i^t)^k \in \mathbb{R}^{F/K}$, $k = 1, 2, \ldots, K$. Then, K sets of independent attention coefficients are calculated, and their features are concatenated afterwards to maintain the dimensions of output vectors consistent with input vectors for each GAT layer. This process yields the following feature vectors:

$$\tilde{\boldsymbol{h}}_i^{t+1} = \Big\|_{k=1}^{K} \sigma\Big(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^{t+1,\,k}\,(\boldsymbol{W}^{t+1}\boldsymbol{h}_j^t)^k\Big), \tag{6}$$

where σ is the rectified linear unit (ReLU) activation function, $\|$ represents concatenation, $\alpha_{ij}^{t+1,\,k}$ are normalized attention coefficients computed by the k-th attention mechanism $f^{t+1,\,k}$. In this manner, we acquire a technique for updating the node feature vectors that bears resemblance to the CNNs and coarse-grained process employed

in renormalization group theory.[23,53,54] After M rounds of message passing, two final latent representations $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are obtained by the decoder which processes $\boldsymbol{H}^M$ with MLPs.

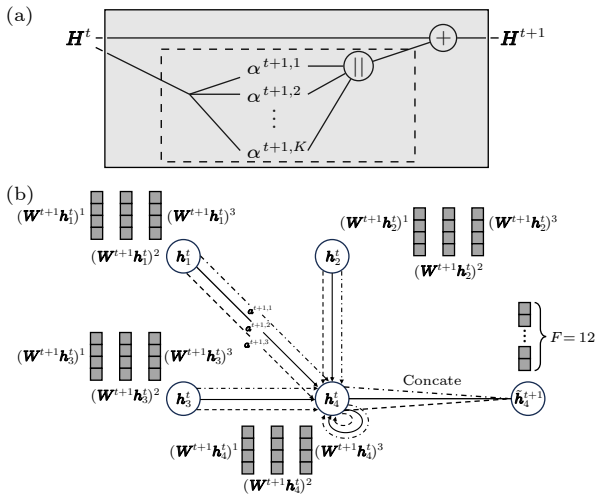

Fig. 4. (a) A schematic for the implementation of message passing functions; (b) an illustration of multi-head attention mechanism for the dash line circled area of (a) (with $K = 3$ heads and $F = 12$) by node 4 on its neighborhood. Different arrow styles denote independent attention heads. The attention mechanism $f^{t+1,\,k}((\boldsymbol{W}^{t+1}\boldsymbol{h}_i^t)^k \| (\boldsymbol{W}^{t+1}\boldsymbol{h}_j^t)^k)$ computes attention coefficients $\alpha_{ij}^{t+1,\,k}$, parameterized by weight vectors $\boldsymbol{a}^{t+1,\,k} \in \mathbb{R}^{2F/K}$, applying LeakyReLU activations. The aggregated features from each head are concatenated to obtain $\tilde{\boldsymbol{h}}_4^{t+1} \in \mathbb{R}^F$.

For the output layer, combining GAT with sum-based reductions makes variational ansatz size extensive, i.e., the wavefunction automatically factorizes into a product of individual states if the physical system is separated into non-interacting components.[11] Therefore, to enable system size generalization, we evaluate amplitudes by the modulus of the product of all entries in the $\boldsymbol{Z}_1$, which is achieved by using a final activation function $f(x) = \log(|x|)$, followed by summing all entries.[39] For phases of wavefunctions, we obtain scalar outputs by sum-pooling the $\boldsymbol{Z}_2$. These traits improve generalization, and allow us to transfer the model to problems with larger sizes.

Based on the preceding discussion, the computational complexity of GAT layers is $O(|V|\times F\times F)+O(|E|\times F)$,[35] where both $|V|$ and $|E|$ increase as $O(N)$ for common graphs. F is the dimension of node features $\boldsymbol{h}_i^t$ and does not increase with N. Given a fixed value of F, the number of parameters in GAT remains constant regardless of the system size N. The overall computational complexity of our model is therefore $O(N)$, meeting all the requirements for the scalability. Compared to RBM and RNN-based NQS with the number of parameters or computational overhead grows quadratically with N, our GAT-based method is easier to be used for large systems.

For test problems, we use similar architectures to emphasize the generality of the proposed method. The node functions of the encoder and decoder are implemented as MLPs with 3 and 2 hidden layers, respectively, with ReLU activations and layer normalizations.[55] The node vectors have sizes of 32 and 48 for nonfrustrated and frustrated systems, respectively, and their sizes are maintained throughout the encode-process-decode process. After each update, the layer normalization is applied to feature vectors. We use 6 message passing steps with residual connections in the processor, and apply the most appropriate number of heads for different systems. These models have 11810 and 25394 variational parameters (equivalent respectively to 0.047 M and 0.10 M Bytes), which enable us to optimize trainable parameters by SR.

Although our GAT method shares common features with the graph convolutional network (GCN)[36] and other GNN-based[40] approaches, there are significant differences. The existing GCN-based model is known to have limited expressiveness due to the problem of over-smoothing,[52] and cannot be applied to frustrated systems due to its network structure.[36] Meanwhile, the GNN model with message passing process based on both node and edge features[40] requires 0.86 M trainable parameters for the problems tested here. The numbers of training parameters suggest that the self-attention mechanism of GAT is more conducive than the message passing GNN to encode spin correlations, as also observed in Refs. [31,50]. Moreover, the excessive number of parameters makes GNN difficult to use the advanced SR method.[31]

Test Problems. To evaluate the performance of our model, we deploy it on a set of Heisenberg Hamiltonians:

$$\hat{H}_{J_1-J_2} = J_1 \sum_{\langle i,j\rangle} \hat{\boldsymbol{\sigma}}_i \cdot \hat{\boldsymbol{\sigma}}_j + J_2 \sum_{\langle\langle i,j\rangle\rangle} \hat{\boldsymbol{\sigma}}_i \cdot \hat{\boldsymbol{\sigma}}_j, \tag{7}$$

where $\hat{\boldsymbol{\sigma}}_i = (\hat{\sigma}_i^x, \hat{\sigma}_i^y, \hat{\sigma}_i^z)$ is the vector of Pauli operators on spin i; $\langle i, j\rangle$ and $\langle\langle i, j\rangle\rangle$ correspond to indices of nearest and next nearest neighbors, respectively. We consider lattices of various sizes and interaction strengths with periodic boundary conditions, and without loss of generality, take $J_1 = 1$ in all calculations of this work. The ground states of such Hamiltonians are strongly influenced by the underlying geometry and the relative interaction strength J_2/J_1 between next nearest and nearest neighbors. Following Kochkov *et al.*[40] and Fu *et al.*,[37] our test suite includes four routinely studied geometric variations: square, honeycomb, triangular, and kagome lattices. Each of these Hamiltonians contains multiple phases, including ordered and disordered domains. Disordered regimes of the square and honeycomb lattices are realized by J_2/J_1 of 0.5[56–59] and 0.2,[60,61] respectively. The addition of geometric frustration in triangular and kagome lattices results in even larger quantum fluctuations and more challenging ground states to learn.

The square lattice J_1–J_2 model is one of the most important frustrated magnet models in the early days of high-T_c research, and it has attracted intense research interest, both theoretically and experimentally. In the vicinity of the classical boundary $J_2/J_1 = 0.5$, the model is frustrated and the exact phase diagram is subject to ongoing debate.[56–58,62] Still, the square lattice J_1–J_2 model is widely recognized as the benchmark for various approximate methods.[32] Therefore, the energy and order parameter results of the GAT ansatz on square lattices are first used to verify the feasibility of our model. Subsequently, we extend this approach to non-square lattices.

Results and Discussion. During training, the model parameters are optimized by SR, with a number of samples

for each SR step at 6000, and the post-selection method [63] is adopted in order to obtain reliable estimates of the energy. The optimization generally converges within 30000 steps. During testing, the energy and order parameters are estimated using 10^6 configurations sampled from equilibrated Markov chains. The simulations are implemented using the jVMC library. [33]

The performance of our models is evaluated on the aforementioned frequently studied Heisenberg Hamiltonians. Our models are trained for each Hamiltonian specification and their performances are compared to machine learning-based and other methods in literature.

Firstly, the GAT models are assessed on test problems with a system size of 36/32, which can be precisely solved using the Lanczos algorithm. According to Li *et al.*, [64] when learning ground states of frustrated models, direct training can be difficult for the networks to converge towards optimal results, while transfer learning can greatly speed up the optimization processes and obtain better results. Thus, after a model is optimized for a lattice, its parameters are served as the initial parameters for the same model with larger N. As an example, these optimization procedures on square lattices at $J_2/J_1 = 0.5$ are shown in Fig. 5. This confirms that the representation learned by GAT is local, and our model can be applied to systems of different sizes with retraining.

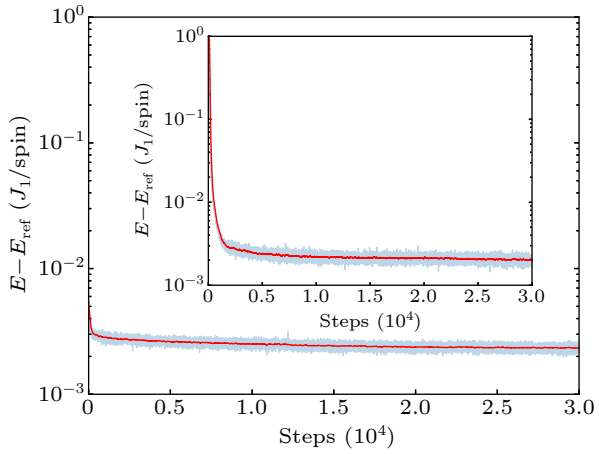

Fig. 5. Optimization of the GAT-based NQS for the J_1–J_2 Heisenberg model at $J_2/J_1 = 0.5$ on the 10×10 square lattice. We initialize the parameters with values from 6×6 lattice, whose optimization process is shown in the inset. The reference energies $E_{\rm ref}$ are obtained by RBM combined with pair-product states (PP), referred as RBM+PP [30] and exact diagonalization [65] for 10×10 and 6×6 lattices, respectively. The energy expectation is averaged over the last 100 SR steps, as indicated by the solid line.

The ground state energies for the nonfrustrated, i.e., $J_2 = 0$, models on square lattices can be obtained precisely by the stochastic series expansion method. [66] Here we find a variational energy of -0.67139 per spin. Our result is only about 1.6×10^{-4} higher than the best result of -0.671549, and better than several other neural network-based wavefunction methods such as CNNs [24,39] and RBM-enhanced GWF. [28] Part of the reason for us to get such a precise result is that the sublattice encoding implies the Marshall sign rule, which is the exact sign structure for the ground states of the antiferromagnetic Heisenberg models on square lattices. [67] In Ref. [68], aCNN also obtains cutting-edge results on nonfrustrated models by using the Marshall sign rule directly.

We also study the antiferromagnetic Heisenberg models on $L = 16$ and $L = 24$ square lattices. Results are shown in Table 1 and compared with those of VMC, QMC and aCNN. As seen in Table 1, our GAT-based results are about 2×10^{-4} higher than those gained by QMC [69] for all lattice sizes. Notice that, while our result is 4×10^{-5} higher than aCNN on the 10×10 lattice, it is lower than those of aCNN by 6×10^{-5} and 2×10^{-4} for $L = 16$ and $L = 24$. That is, the GAT method is increasingly better than aCNN with the increased lattice size. The results demonstrate that our GAT-based NQS exhibits superior performance in solving quantum many-body problems on large systems compared to traditional and some CNN-based methods, confirming the favorable expressivity and scalability of GAT.

Table 1. Ground state energies of the antiferromagnetic Heisenberg models on square lattices obtained in this work (GAT) and literature (with error bars in parenthesis). The data for VMC, QMC and aCNN are referenced from Refs. [66,68,69].

Method	$L = 10$	$L = 16$	$L = 24$
VMC [69]	–	−0.66965	−0.66922
QMC [66,69]	−0.671549(4)	−0.66998	−0.66961
aCNN [68]	−0.671431(2)	−0.669716(1)	−0.669237(2)
GAT	−0.671390(3)	−0.669781(2)	−0.669433(1)

Table 2. Comparison of ground state energies of the spin-1/2 J_1–J_2 model on a 10×10 square lattice, obtained in this work (GAT) and the state-of-the-art literature results. The "best" results are obtained with numerically exact stochastic series expansion for $J_2 = 0$ [66] and RBM+PP, [30] transformer [31] for $J_2/J_1 = 0.5$.

Method	$J_2 = 0$	$J_2/J_1 = 0.5$
GWF [27]	−0.66935(1)	−0.49439(1)
GWF+RBM [28]	−0.67111(2)	−0.49575(3)
CNN1 [24]	−0.67135(1)	−0.49516(1)
CNN2 [39]	−0.671275(5)	−0.494757(12)
aCNN [68]	−0.671431(2)	−0.495627(6)
Best [30,31,66]	−0.671549(4)	−0.497629(1)
		−0.497634(1)
GAT	−0.671390(3)	−0.49548(1)

Our approach is then used to investigate the fully frustrated phase of the J_1–J_2 model at $J_2/J_1 = 0.5$. For the square lattice with $L = 6$, the exact diagonalization gives the result of -0.50381. [65] Bukov *et al.* [63] identify an energy error limit that appears to be hard to overcome for NQS, and multiple works [24,28,39,63] have indeed all yielded the energy error on the order of 2×10^{-3}. Our result of -0.50189 is also comparable to the NQS values reported in literature.

For the 10×10 square lattice, the energies obtained for $J_2 = 0$ and 0.5, together with relevant benchmarks, are summarized in Table 2. Again, it demonstrates that our results are comparable to the CNN models with the number of parameters within $O(10^4)$. [24,39,68] Compared to the best results, our result for $J_2/J_1 = 0.5$ is slightly

worse than that for $J_2 = 0$, primarily due to the deviation of the former ground state sign structure from the Marshall sign rule.

The best ground state energies for $J_2/J_1 = 0.5$ are obtained by the methods of RBM+PP [30] and transformer wavefunction. [31] Our variational energy of -0.49548 per spin is 2.2×10^{-3} higher than the best results of -0.497629 and -0.497634. RBM+PP [30] is a hybrid model combining neural networks with a standard variational model that incorporates assumptions about the underlying physics. Normally, methods with physics-based assumptions may provide very exact results with relatively few parameters. [23,27,30] However, their performances are highly dependent on the availability of strong physical priors that severely limits their applicability. In comparison, the GAT method is not biased by *a priori* assumptions about the physics and applicable to general systems. Meanwhile, the GAT approach also allows for the use of physical priors as PP [30] or symmetry-enhanced backflow wavefunctions [70] to further improve its accuracy. The recently proposed transformer wavefunction, [31] though finding an energy that is 5×10^{-6} better than RBM+PP using about ten times more parameters of our GAT model, has two problems. The transformer has a computational complexity of $O(N^2)$. [34] Therefore, it is more difficult than the GAT approach to use in large systems. The application of the transformer wavefunction on non-square lattices has not been realized yet.

As an added test, our GAT model is employed to evaluate the antiferromagnetic order parameter defined in Refs. [23,24,39],

$$S^2(\boldsymbol{q}) = \frac{1}{N(N+2)} \sum_{i,j} \langle \hat{\boldsymbol{\sigma}}_i \cdot \hat{\boldsymbol{\sigma}}_j \rangle e^{i\boldsymbol{q}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)}, \tag{8}$$

where $\boldsymbol{q} = (\pi, \pi)$ and $\boldsymbol{q} = (\pi, 0)$ correspond to the structure factors of Néel order and stripe order, respectively. Results are listed in Table 3. For $J_2 = 0$, we find $S^2(\pi, \pi) = 0.1568$ and $S^2(\pi, 0) = 0.00494$, indicating that the system is in Néel order with $\boldsymbol{q} = (\pi, \pi)$. For $J_2/J_1 = 0.5$, we get $S^2(\pi, \pi) = 0.0634$ and $S^2(\pi, 0) = 0.00538$. All the results are close to those given in Refs. [23,24,39].

Table 3. Comparison of antiferromagnetic order parameters of the spin-1/2 J_1–J_2 model on a 10×10 square lattice, obtained in this work (GAT) and literature results given by CNN and hybrid wavefunction combining CNN and projected entangled pair states (PEPS). [23,39]

Method	$J_2 = 0$		$J_2/J_1 = 0.5$	
	$S^2(\pi, 0)$	$S^2(\pi, \pi)$	$S^2(\pi, 0)$	$S^2(\pi, \pi)$
CNN2 [39]	0.00498(5)	0.1571(2)	0.00521(7)	0.0633(2)
PEPS+CNN [23]	0.00496	0.15665	0.00633	0.05880
GAT	0.00494(3)	0.1568(1)	0.00538(3)	0.0634(1)

There are relatively few literature results about solving J_1–J_2 models on non-square lattices using neural network-based methods. Table 4 summarizes the comparisons of our GAT ansatz results with the state-of-the-art literature results, such as those by GNN and ResNet. [40] For small systems ($N = 32$, 36), the reference energies are obtained by exact diagonalization, while for large systems ($N = 98$, 108), they are found with approximation approaches, such as DMRG and extrapolation method. [37]

Table 4 shows that the wavefunctions learned by the proposed GAT ansatz consistently give energies close to the reference results on both small and large systems, and the GAT performs relatively well on large lattices. For honeycomb lattices, GAT outperforms ResNet and is comparable to GNN in all cases. In fact, the GAT results are very close to the exact ones for small honeycomb systems and better than the references to large honeycomb systems. For triangular lattices, GAT outperforms ResNet, but is comparable with GNN. For kagome lattices, GAT, though comparable to ResNet, is slightly inferior to GNN, probably because these systems have greater frustration, leading to even larger quantum fluctuations and harder to be evaluated by the present small model. Overall, the good performance shown above proves that the proposed GAT is capable of accurately representing the ground states of quantum many-body systems.

Table 4. Estimated energy per site of the learned wavefunction. Ground states of J_1–J_2 models on three kinds of lattice with various sizes and interaction strengths are used for comparison. The GAT results below or close to GNN are denoted in bold. The data for comparison are referenced from Refs. [37,40].

Lattice	Size	J_2	GNN [40]	ResNet [40]	GAT	Reference energy [37]
Honeycomb	32	0	-0.551 [37]	–	**$-0.551640(3)$**	-0.5517
		0.2	$-0.4563(6)$	$-0.4561(1)$	**$-0.456310(8)$**	-0.4567
	98	0	–	–	$-0.545911(2)$	-0.5448
		0.2	$-0.4528(2)$	$-0.4496(4)$	**$-0.453602(4)$**	-0.4527
Triangular	36	0	$-0.5571(5)$ [37]	–	**$-0.55851(2)$**	-0.5603734
		0.125	$-0.512(2)$	$-0.5007(5)$	$-0.50944(2)$	-0.515564
	108	0	$-0.5508(8)$	$-0.5451(2)$	**$-0.55083(1)$**	-0.551
		0.125	$-0.500(9)$	$-0.488(9)$	**$-0.50374(2)$**	-0.5126
Kagome	36	0	$-0.434(1)$	$-0.433(0)$	$-0.42987(2)$	-0.43837653
	108	0	$-0.4302(1)$	$-0.424(9)$	$-0.42786(1)$	-0.4386

Considering the data in Table 4, GNN is seen to produce relatively accurate results. However, it merges edge features into nodes in each iteration, and all the features may tangle up together and make the network complicated and confusing. Moreover, the excessive number of parameters [31] makes GNN difficult to use advanced SR optimization methods, and may hinder the GNN's ability to combine with physics-based methods to find more accurate results. Compared to 0.86 M trainable parameters in GNN, our GAT models only have about 0.047 M and 0.10 M parameters. Though with much fewer parameters than GNN, GAT can achieve performance on par or

better than that. The good performance is mainly due to the use of attention mechanisms to capture correlations between spins, which enhances the expressivity of the ansatz. Meanwhile, it is also due to the utilization of two final latent representations that somewhat decouples the variational parameters responsible for the highly oscillatory phases and slower varying magnitudes. Moreover, considering the small number of trainable parameters in the current GAT models, there are rooms for improvement of the results. For example, the results of GAT for kagome lattices can be improved by adding more trainable parameters and using the recently proposed minSR optimization method.[71]

Our GAT model is also applicable to spin models other than the Heisenberg model. For example, Table 5 shows there is a good agreement between our method and the best literature results about the ground state energies of the transverse field Ising model (TFIM),

$$\hat{H}_{\mathrm{TFIM}} = -\sum_{\langle i,j\rangle} \hat{\sigma}_i^z \hat{\sigma}_j^z - \lambda \sum_i \hat{\sigma}_i^x, \tag{9}$$

with transverse field $\lambda = 3.05$, which is close to the phase transition point of the TFIMs on square lattices. Numerical experiments on various systems as discussed above demonstrate that our model has strong expressivity, and characterizes quite accurately the subspace of Hilbert space where wavefunctions reside in.

Table 5. Ground state energies of the TFIMs on square lattices obtained here (GAT) and by aCNN[68] and the extrapolated results of Fully Augmented Tree Tensor Network (FATTN)[72] and Fully Augmented Matrix Product State (FAMPS).[73]

Method	$L = 8$	$L = 16$
FATTN[72]	–	−3.23956(2)
FAMPS[73]	−3.24165	–
aCNN[68]	−3.241642(2)	−3.239568(2)
GAT	−3.241650(1)	−3.239580(1)

In summary, we have proposed a GAT-based variational ansatz with an encode-process-decode architecture to learn the distributed representation of the wavefunction that respects the underlying symmetries of the problem. The GAT ansatz employs a self-attention mechanism to capture spin-spin correlations, thereby enhancing its expressivity. The computational complexity of the GAT ansatz grows linearly with the number of lattice sites and can be applied to large systems without a significant loss of accuracy.

Numerical experiments show that the GAT model with a small number of trainable parameters can accurately capture the ground state physics for a variety of problems without relying on prior knowledge about the target structure. The GAT method provides state-of-the-art results for Heisenberg models across different lattices and interaction strengths, including challenging problems of frustrated domains that are difficult to solve due to complicated sign structures. It also provides excellent results for the TFIMs. These tests demonstrate that the GAT ansatz is efficient and versatile for solving the ground states of quantum spin models. The GAT-based techniques hold promise of being a useful tool for studying quantum many-body systems and providing deeper insights into and implicit modeling of exponentially sized objects.

The GAT algorithm developed here has been released as open source code in the Github repository at https://github.com/GATwf/GAT.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 12374017 and 12074362) and the Innovation Program for Quantum Science and Technology (Grant No. 2021ZD0303303). The computing time of the Supercomputing Center of the University of Science and Technology of China is gratefully acknowledged.

References

[1] White S R 1992 *Phys. Rev. Lett.* **69** 2863
[2] Schollwöck U 2011 *Ann. Phys.* **326** 96
[3] Orús R 2014 *Ann. Phys.* **349** 117
[4] Biamonte J and Bergholm V 2017 arXiv:1708.00006v1 [quant-ph]
[5] Choo K, Carleo G, Regnault N, and Neupert T 2018 *Phys. Rev. Lett.* **121** 167204
[6] Ceperley D and Alder B 1986 *Science* **231** 555
[7] Foulkes W M C, Mitas L, Needs R J, and Rajagopal G 2001 *Rev. Mod. Phys.* **73** 33
[8] Carlson J, Gandolfi S, Pederiva F, Pieper S C, Schiavilla R, Schmidt K E, and Wiringa R B 2015 *Rev. Mod. Phys.* **87** 1067
[9] Troyer M and Wiese U J 2005 *Phys. Rev. Lett.* **94** 170201
[10] McMillan W L 1965 *Phys. Rev.* **138** A442
[11] Becca F and Sorella S 2017 *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge: Cambridge University Press)
[12] Hornik K, Stinchcombe M, and White H 1989 *Neural Netw.* **2** 359
[13] Cybenko G 1989 *Math. Control Signals Systems* **2** 303
[14] Hornik K 1991 *Neural Netw.* **4** 251
[15] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L, and Zdeborová L 2019 *Rev. Mod. Phys.* **91** 045002
[16] Carrasquilla J 2020 *Adv. Phys.: X* **5** 1797528
[17] Ma X R, Tu Z C, and Ran S J 2021 *Chin. Phys. Lett.* **38** 110301
[18] Kong J G, Li Q X, Li J, Liu Y, and Zhu J J 2022 *Chin. Phys. Lett.* **39** 067503
[19] Carleo G, Troyer M 2017 *Science* **355** 602
[20] Gao X and Duan L M 2017 *Nat. Commun.* **8** 662
[21] Saito H and Kato M 2018 *J. Phys. Soc. Jpn.* **87** 014001
[22] Liang X, Liu W Y, Lin P Z, Guo G C, Zhang Y S, and He L 2018 *Phys. Rev. B* **98** 104426
[23] Liang X, Dong S J, and He L 2021 *Phys. Rev. B* **103** 035138
[24] Choo K, Neupert T, and Carleo G 2019 *Phys. Rev. B* **100** 125124
[25] Hibat-Allah M, Ganahl M, Hayward L E, Melko R G, and Carrasquilla J 2020 *Phys. Rev. Research* **2** 023358
[26] Sfondrini A, Cerrillo J, Schuch N, and Cirac J I 2010 *Phys. Rev. B* **81** 214426
[27] Hu W J, Becca F, Parola A, and Sorella S 2013 *Phys. Rev. B* **88** 060402
[28] Ferrari F, Becca F, and Carrasquilla J 2019 *Phys. Rev. B* **100** 125131
[29] Gong S S, Zhu W, Sheng D N, Motrunich O I, and Fisher M P A 2014 *Phys. Rev. Lett.* **113** 027201
[30] Nomura Y and Imada M 2021 *Phys. Rev. X* **11** 031034

[31] Rende R, Viteritti L L, Bardone L, Becca F, and Goldt S 2023 arXiv:2310.05715v1 [cond-mat.str-el]
[32] Reh M, Schmitt M, and Gärttner M 2023 *Phys. Rev. B* **107** 195115)
[33] Schmitt M and Reh M 2022 *SciPost Phys. Codebases* p 2
[34] Batatia I, Schaaf L L, Chen H, Csányi G, Ortner C, and Faber F A 2023 arXiv:2310.10434v1 [stat.ML]
[35] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, and Bengio Y 2018 arXiv:1710.10903v3 [stat.ML]
[36] Yang L, Hu W, and Li L 2020 arXiv:2011.12453v1 [cond-mat]
[37] Fu C, Zhang X, Zhang H, Ling H, Xu S, and Ji S 2022 arXiv:2206.07370v1 [quant-ph]
[38] Park C Y and Kastoryano M J 2020 *Phys. Rev. Research* **2** 023232
[39] Szabó A and Castelnovo C 2020 *Phys. Rev. Res.* **2** 033075
[40] Kochkov D, Pfaff T, Sanchez-Gonzalez A, Battaglia P, and Clark B K 2021 arXiv:2110.06390v1 [quant-ph]
[41] Messio L, Lhuillier C, and Misguich G 2011 *Phys. Rev. B* **83** 184401
[42] Wen X G 2002 *Phys. Rev. B* **65** 165113
[43] Battaglia P W, Hamrick J B, Bapst V *et al.* 2018 arXiv:1806.01261v3 [cs.LG]
[44] Bishop C M 2006 *Pattern Recognition And Machine Learning* (New York: Springer)
[45] He K M, Zhang X Y, Ren S Q, and Sun J 2016 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, USA, June 26–July 1, 2016) p 770
[46] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, and Polosukhin I 2017 arXiv:1706.03762 [cs.CL]
[47] Devlin J, Chang M W, Lee K, and Toutanova K 2019 arXiv:1810.04805v2 [cs.CL]
[48] Goodall R E A and Lee A A 2020 *Nat. Commun.* **11** 6280
[49] Xie T, France-Lanord A, Wang Y, Shao-Horn Y, and Grossman J C 2019 *Nat. Commun.* **10** 2667
[50] Viteritti L L, Rende R, and Becca F 2023 *Phys. Rev. Lett.* **130** 236401
[51] von Glehn I, Spencer J S, and Pfau D 2023 arXiv:2211.13672v2 [physics.chem-ph]
[52] Cai C and Wang Y S 2020 arXiv:2006.13318v1 [cs.LG]
[53] Kadanoff L P 2000 *Statistical Physics: Statics, Dynamics and Renormalization* (Singapore: World Scientific)
[54] Yang X C, Xie Z Y, and Yang X T 2023 *Chin. Phys. Lett.* **40** 020501
[55] Ba J L, Kiros J R, and Hinton G E 2016 arXiv:1607.06450v1 [stat.ML]
[56] Dagotto E and Moreo A 1989 *Phys. Rev. Lett.* **63** 2148
[57] Zhitomirsky M E and Ueda K 1996 *Phys. Rev. B* **54** 9007
[58] Capriotti L, Becca F, Parola A, and Sorella S 2001 *Phys. Rev. Lett.* **87** 097201
[59] Jiang H C, Yao H, and Balents L 2012 *Phys. Rev. B* **86** 024424
[60] Albuquerque A F, Schwandt D, Hetényi B, Capponi S, Mambrini M, and Läuchli A M 2011 *Phys. Rev. B* **84** 024406
[61] Mosadeq H, Shahbazi F, and Jafari S A 2011 *J. Phys.: Condens. Matter* **23** 226006
[62] Liu W Y, Gong S S, Li Y B, Poilblanc D, Chen W Q, and Gu Z C 2022 *Sci. Bull.* **67** 1034
[63] Bukov M, Schmitt M, and Dupont M 2021 *SciPost Phys.* **10** 147
[64] Li M F, Chen J S, Xiao Q, Wang F, Jiang Q C, Zhao X C, Lin R F, An H, Liang X, and He L X 2022 *IEEE Trans. Parallel Distrib. Syst.* **33** 2846
[65] Schulz H J, Ziman T A L, and Poilblanc D 1996 *Journal de Physique I* **6** 675
[66] Mezzacapo F, Schuch N, Boninsegni M, and Cirac J I 2009 *New J. Phys.* **11** 083026
[67] Marshall W and Peierls R E 1955 *Proc. R. Soc. Lond. A* **232** 48
[68] Wang J Q, He R Q, and Lu Z Y 2023 arXiv:2308.09664v1 [cond-mat.str-el]
[69] Sandvik A W and Evertz H G 2010 *Phys. Rev. B* **82** 024407
[70] Luo D and Clark B K 2019 *Phys. Rev. Lett.* **122** 226401
[71] Chen A and Heyl M 2023 arXiv:2302.01941v2 [cond-mat.dis-nn]
[72] Qian X J and Qin M P 2022 *Phys. Rev. B* **105** 205102
[73] Qian X J and Qin M P 2023 *Chin. Phys. Lett.* **40** 057102