UNIVERSITÀ DEGLI STUDI DI TRIESTE

Transformer Variational Wave Functions for Frustrated Quantum Spin Systems

This is the peer reviewed (post-print) version of the following article:

Original
Transformer Variational Wave Functions for Frustrated Quantum Spin Systems / Viteritti, L.L., Rende, R., Becca, F.. - In: PHYSICAL REVIEW LETTERS. - ISSN 0031-9007. - 130/2023:23(2023), pp. 236401.1-236401.6. [10.1103/physrevlett.130.236401]

Availability:
This version is available at: 11368/3075239 since: 2024-05-23T11:14:26Z

Publisher:

Published
DOI:10.1103/physrevlett.130.236401

Terms of use:

Some rights reserved. More details in the PostPrint rights section below.

License:
Digital Rights Management not defined

Publisher copyright

-.-

(Article begins on next page)

21 October 2026

arTs Università degli Studi di Trieste
Archivio della ricerca – postprint

Transformer Variational Wave Functions for Frustrated Quantum Spin Systems

Luciano Loris Viteritti,[1,*] Riccardo Rende,[2,†] and Federico Becca[1]
[1]*Dipartimento di Fisica, Università di Trieste, Strada Costiera 11, I-34151 Trieste, Italy*
[2]*International School for Advanced Studies (SISSA), Via Bonomea 265, I-34136 Trieste, Italy*

The transformer architecture has become the state-of-art model for natural language processing tasks and, more recently, also for computer vision tasks, thus defining the vision transformer (ViT) architecture. The key feature is the ability to describe long-range correlations among the elements of the input sequences, through the so-called self-attention mechanism. Here, we propose an adaptation of the ViT architecture with complex parameters to define a new class of variational neural-network states for quantum many-body systems, the ViT wave function. We apply this idea to the one-dimensional J_1-J_2 Heisenberg model, demonstrating that a relatively simple parametrization gets excellent results for both gapped and gapless phases. In this case, excellent accuracies are obtained by a relatively shallow architecture, with a single layer of self-attention, thus largely simplifying the original architecture. Still, the optimization of a deeper structure is possible and can be used for more challenging models, most notably highly frustrated systems in two dimensions. The success of the ViT wave function relies on mixing both local and global operations, thus enabling the study of large systems with high accuracy.

Introduction.—Variational approaches for studying quantum many-body systems have proved fundamental for understanding the properties of extremely complicated physical systems, famous examples being the Bardeen-Cooper-Schrieffer state [1] and Laughlin [2] wave functions to explain superconductivity and fractional quantum Hall effect, respectively. Given the exponential growth of the many-body Hilbert space, a compact representation of the ground state, encoding the correct physical properties, is a highly nontrivial task for strongly interacting systems. Recently, a class of wave functions, based on neural networks, has been introduced and developed [3,4]. Starting from restricted Boltzmann machines (RBMs) [3], which are the simplest neural-network *Ansatz* (namely only one fully connected hidden layer), numerous studies have been carried out testing different types of architectures; examples include convolutional-neural networks (CNNs) [5–8], recurrent-neural networks (RNNs) [9,10], and autoregressive-neural networks [11,12], but also combinations of neural networks with standard variational wave functions (e.g., Gutzwiller-projected fermionic ones) [13,14].

In the last few years, the transformer architecture [15] has become the state-of-art choice in natural-language processing tasks. Its key feature is the ability to model relationships among all elements of an input sequence (regardless of their positions), by efficiently *transforming* input sequences into abstract representations. Inspired by successes in natural-language processing, very small modifications led to the ViT [16], which has been applied to image classification tasks, achieving competitive results with respect to state-of-art deep CNNs, while being much more efficient than them. Within many-body problems, transformer networks have recently been employed in the context of lattice gauge theories [11], to perform quantum tomography in presence of noise [17], and for real- and imaginary-time evolutions of quantum systems [18].

In this Letter, we demonstrate that the ViT architecture can be adapted to define a new class of neural-network quantum states, here dubbed as ViT wave functions. We apply our *Ansatz* to the one-dimensional J_1-J_2 Heisenberg model, whose Hamiltonian is defined by

$$\hat{H} = J_1 \sum_R \hat{\boldsymbol{S}}_R \cdot \hat{\boldsymbol{S}}_{R+1} + J_2 \sum_R \hat{\boldsymbol{S}}_R \cdot \hat{\boldsymbol{S}}_{R+2} \tag{1}$$

where $\hat{\boldsymbol{S}}_R = (S_R^x, S_R^y, S_R^z)$ is the $S = 1/2$ spin operator at site R and $J_1 > 0$ and $J_2 \geq 0$ are nearest- and next-nearest-neighbor antiferromagnetic couplings, respectively. Its phase diagram is well established by analytical and numerical studies [19]. For small values of J_2/J_1, the ground state has power-law spin-spin correlations, and the excitation spectrum is gapless; for large values of J_2/J_1, the ground state is twofold degenerate, leading to long-range dimer order (but exponentially decaying spin-spin correlations), and the spectrum is fully gapped. These two phases are separated by a critical point at $(J_2/J_1)_c = 0.241167 \pm 0.000005$ [20,21]. Interestingly, for $J_2/J_1 > 0.5$, incommensurate (but short-range) spin-spin correlations have been found, whereas dimer–dimer correlations are always commensurate. In the following, we assess the ground-state properties of the J_1-J_2 model on finite clusters, imposing periodic boundary conditions.

From the numerical perspective, density-matrix renormalization group (DMRG) [22] or its modern variations based upon tensor networks *Ansätze* [23] represent one of the few approaches that can accurately assess the ground-state properties of frustrated systems in one dimension, as the J_1-J_2 model of Eq. (1). In fact, the main limitation to the use of quantum Monte Carlo techniques [24] relies on the unknown sign structure of the ground-state wave function, which prevents one from performing unbiased projection techniques (except for $J_2 = 0$, where the so-called Marshall sign rule applies [25]). The nontrivial sign structure also represents an obstacle to the definition of accurate variational wave functions. For example, Gutzwiller-projected fermionic states [26] have a limited power to reproduce the correct signs of the ground state for $J_2/J_1 > 0.5$ [27]. By contrast, RBM states are able to reach an excellent accuracy; however, they suffer from poor scaling behavior, due to their *fully-connected* structure in which a single hidden layer is connected to all physical degrees of freedom [27]. This fact limits the applicability of RBMs to relatively small clusters. In this respect, CNN wave functions have been introduced to deal with *local* structures, and deep architectures are necessary to build long-range correlations, thus introducing severe problems in the optimization procedure (e.g., diverging or vanishing gradients). RNN *Ansätze* have been also considered, which recurrently process inputs of a sequence one by one, implying that they cannot be parallelized; in addition, since not all elements of the network are directly connected, long-range correlations are built from short-range ones, thus making the learning process not straightforward [28].

In order to overcome these problems, we propose a simplified version of the standard ViT architecture. The main advantage of this *Ansatz* lies in the possibility of mixing both local and global structures, thus limiting the number of variational parameters and simplifying the learning process (see below). We emphasize that a complex parametrization is adopted without an *a priori* encoding of the sign structure (i.e., no information about the exact signs). In this work, we show that the ViT wave function can reach very high accurate results compared with DMRG calculations, even on large clusters, with fewer than 1 000 parameters and few computational resources compared with other neural-network wave functions. Most importantly, the ViT accuracy can be systematically improved by changing the hyperparameters of the architecture.

Methods.—The fundamental ingredient of a transformer is the *self-attention mechanism*. Given a sequence of N input vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, for each of them three new vectors are computed, $\boldsymbol{q}_i = Q\boldsymbol{x}_i$, $\boldsymbol{k}_i = K\boldsymbol{x}_i$, and $\boldsymbol{v}_i = V\boldsymbol{x}_i$, where Q, K, V are generic rectangular matrices of parameters. The attention vectors are then constructed, $\boldsymbol{A}_i = \sum_{j=1}^{N} \alpha(\boldsymbol{q}_i, \boldsymbol{k}_j)\boldsymbol{v}_j$, where the attention weights $\alpha(\boldsymbol{q}_i, \boldsymbol{k}_j)$ determine how much the jth input vector should contribute to $\boldsymbol{A}_i$, which is the subsequent representation of the ith input. The functional form of these weights can be chosen according to the task [29]. To improve the performance of the model, multihead attention can be considered, where a set of matrices Q^μ, K^μ, and V^μ, with $\mu = 1, \ldots, h$ (with h the *number of heads*) is defined, thus leading to a set of attention vectors $\boldsymbol{A}_i^\mu$. The latter ones are computed in parallel, concatenated together, and linearly combined. Finally, each output vector of the multihead attention is fed separately and identically to a nonlinearity. In general, this whole architecture is replicated n_l times.

Our goal is to use the transformer to parametrize the many-body wave function, in order to map spin configurations of the Hilbert space $\sigma = (\sigma_1, \ldots, \sigma_L)$, with $\sigma_R = 2S_R^z = \pm 1$, to complex numbers $\Psi(\sigma)$. We take inspiration from the ViT [16] introduced for computer vision tasks, where the images are split into patches and these are taken as the input sequence to a transformer. In the same way, starting from a spin configuration $\sigma = (\sigma_1, \ldots, \sigma_L)$, we split it into N patches of b elements: $\boldsymbol{x}_i = (\sigma_{(i-1)b+1}, \ldots, \sigma_{(i-1)b+b})$, for $i = 1, \ldots, N$ (the total number of sites must be a multiple of b). The sequence of these patches is then used to compute the attention vectors. Then, a simplification of the original ViT is considered, taking the attention weights only depending on positions i and j, but not on the actual values of the spins in these patches, thus leading to

$$\boldsymbol{A}_i^\mu = \sum_{j=1}^{N} \alpha_{ij}^\mu \mathrm{V}^\mu \boldsymbol{x}_j, \tag{2}$$

where V^μ is a $r \times b$ matrix with $r = d/h$, and d is the so-called *embedding dimension* that must be a multiple of the number of heads h. This approach is dictated by the fact that the attention weights should mainly depend on the relative positions among groups of spins and not on the actual values of the spins in the patches. This is expected to be true when the patches are far apart and is extended for generic positions i and j. Finally, after the concatenation of the heads, a further linear projection is taken, before the nonlinearity, here chosen as $\log[\cosh(\cdot)]$. This block can be repeated n_l times before applying the output layer in which all the values are summed to obtain the logarithm of the ViT wave function $\Psi_{\mathrm{ViT}}(\sigma)$ (see Fig. 1).

In order to study frustrated quantum spin models with a nonpositive ground state (in the computational basis), we choose all the parameters to be *complex numbers*. Furthermore, a translationally invariant wave function with $k = 0$ can be easily defined by considering the following two steps. First, we adapt the *relative positional encoding* [30] to periodic systems, taking $\alpha_{i,j}^\mu = \alpha_{i-j}^\mu$; as a result, the number of variational parameters for computing the attention vectors [Eq. (2)] is reduced from $O(L^2)$ to $O(L)$. This procedure induces translational invariance between patches. To also include the one within patches, we perform the linear combination

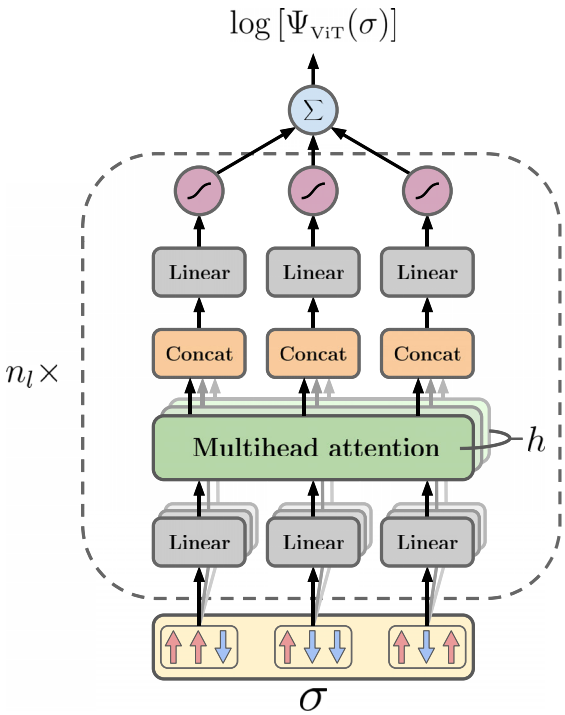

FIG. 1. Scheme of the ViT wave function. The input spin configuration σ is split into patches of size b (which define a set of N vectors of dimension b). Each of them is linearly projected h times with different linear projections to produce N vectors of dimensions $r = d/h$. Then the attention function is applied in parallel, and the h different r dimensional output vectors $\boldsymbol{A}_i^\mu$ are obtained. Then, they are concatenated to a d dimensional vector $\mathrm{Concat}(\boldsymbol{A}_i^1, \ldots, \boldsymbol{A}_i^h)$ and, after another linear projection, the non-linear function $\log[\cosh(\cdot)]$ is applied. This architecture can be replicated and stacked n_l times. The last layer simply sums all the outputs and returns the logarithm of the ViT wave function.

$$\tilde{\Psi}_{\mathrm{ViT}}(\sigma) = \sum_{r=0}^{b-1} \Psi_{\mathrm{ViT}}(T_r \sigma), \tag{3}$$

where T_r is the translation operator. We emphasize that this approach requires a small summation (of b terms), which does not grow with the system size L.

The optimization process of all the complex parameters is obtained by using standard variational Monte Carlo techniques, namely the so-called stochastic reconfiguration approach (see the Supplemental Material [31] for more details). In the following, we mainly take $n_l = 1$, which represents the simplest possible adaptation of the transformer architecture; indeed, even within this drastic assumption, we obtain excellent results in both gapless and gapped phases. At the end, we show the effect of a deeper network with $n_l > 1$. All the simulations are performed by fixing the patch size $b = 4$.

Results.—We start by discussing how the accuracy of the ViT wave function with one layer can be systematically improved by varying its two hyperparameters, i.e., the number of heads h and the ratio $r = d/h$. We consider a cluster with $L = 100$ sites and three different values of the frustration ratio: $J_2/J_1 = 0$ (unfrustrated, gapless), 0.4 (weakly frustrated, gapped), and 0.8 (strongly frustrated, gapped); the reference energy is computed by using the standard DMRG approach (imposing periodic-boundary conditions on the Hamiltonian [33]). In Fig. 2, we show the accuracy of the ground-state energy for the unfrustrated case as a function of d/h fixing the number of heads h, and for the three values of J_2/J_1 when increasing the number of heads h, at fixed ratio d/h. Even though there is a general difficulty in reconstructing the exact sign structure in highly frustrated regimes [27,34–37], we obtain an excellent approximation of the correct energy for all the values of J_2/J_1 that have been considered, e.g., an accuracy $\Delta\varepsilon \lesssim 0.1\%$ for $J_2/J_1 = 0.8$ and $\Delta\varepsilon \approx 0.01\%$ for $J_2/J_1 = 0.4$.

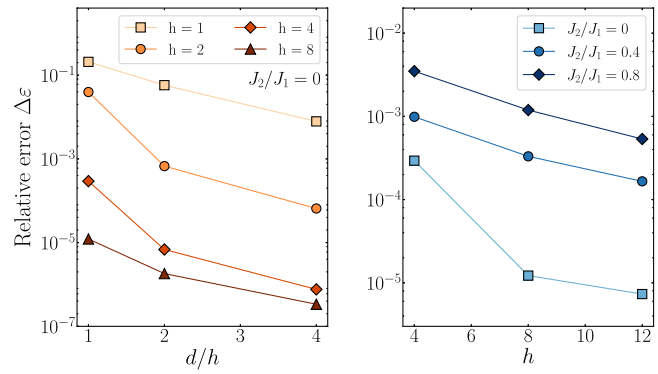

FIG. 2. Relative error $\Delta\varepsilon = |(E_{\mathrm{ViT}} - E_{\mathrm{DMRG}})/E_{\mathrm{DMRG}}|$ of the ViT wave function by varying the hyperparameters of the architecture for a cluster with $L = 100$ sites. Left panel: $\Delta\varepsilon$ as a function of d/h, with a fixed number of heads h, for the unfrustrated case. Right panel: $\Delta\varepsilon$ as a function of the number of heads h, with $d/h = 1$, for different values of frustration ratio. The reference energies are computed by DMRG [33] with a bond dimension up to $\chi = 600$ obtaining $E/J_1 = -0.4432295$ for $J_2/J_1 = 0$, $E/J_1 = -0.3803882$ for $J_2/J_1 = 0.4$, and $E/J_1 = -0.4216664$ for $J_2/J_1 = 0.8$.

Let us now move to the analysis of the correlation functions. From the previous results, we choose $h = 8$ and $d/h = 1$ as a good compromise between accuracy and complexity, for which the network can be trained on $L = 100$ sites in a few hours on ten central processing units (CPUs) or in a few minutes on a graphics processing unit (GPU). The spin-spin correlations are defined as

$$C^{\nu\nu}(r) = \frac{1}{L} \sum_{R=0}^{L-1} \langle \hat{S}_R^\nu \hat{S}_{R+r}^\nu \rangle, \tag{4}$$

where $\nu = x$, y, or z and $\langle \ldots \rangle$ represents the expectation value over the variational quantum state. In particular, we focus on isotropic spin-spin correlations $C(r) = [C^{zz}(r) + C^{xx}(r) + C^{yy}(r)]/3$ and the corresponding structure factor in Fourier space $S(k) = (1/L) \sum_{r=0}^{L-1} e^{ikr} C(r)$. In Fig. 3, we show the results of the real-space correlations $C(r)$ for the

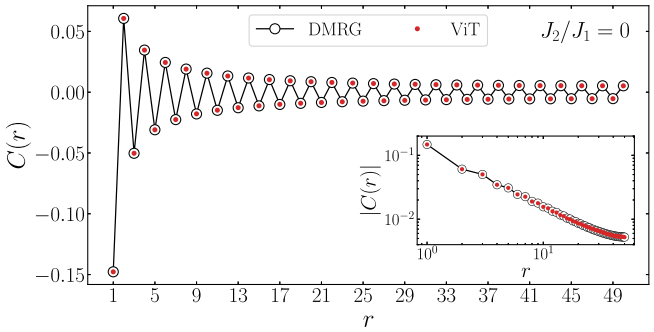

FIG. 3. The isotropic spin-spin correlations in real space $C(r)$ as computed by the ViT wave function (full dots) for the unfrustrated Heisenberg model ($J_2/J_1 = 0$) on a cluster with $L = 100$ sites. The DMRG results are also shown for comparison (empty circles). Inset: log-log plot of the same correlation function.

unfrustrated Heisenberg model ($J_2/J_1 = 0$) on a cluster with $L = 100$ sites, comparing them to the DMRG outcomes (with periodic-boundary conditions). Remarkably, the ViT *Ansatz* is able to match the DMRG calculations at all distances, demonstrating that the global structure of the multihead attention layer is able to build the algebraic long-range tail.

The high flexibility of the ViT state is also demonstrated by considering the three different regimes, with commensurate [i.e., $S(k)$ peaked at $k = \pi$] or incommensurate [i.e., $S(k)$ peaked at $k \neq \pi$] correlations; see Fig. 4.

The gapped phase is characterized by a finite dimer order (implied by the twofold degeneracy of the ground state, in the thermodynamic limit). On any finite system, there is an exponentially small gap between the two states, with $k = 0$ and $k = \pi$, and the insurgence can be detected from the connected dimer-dimer correlations:

$$D(r) = \frac{1}{L}\sum_{R=0}^{L-1}\langle \hat{S}^z_R \hat{S}^z_{R+1} \hat{S}^z_{R+r} \hat{S}^z_{R+r+1}\rangle - [C^{zz}(r=1)]^2 \qquad (5)$$

where $C^{zz}(r = 1)$ is the z component of the spin-spin correlation function at distance $r = 1$ defined in Eq. (4). Notice that this definition considers only the z component of the spin operators [38]. In Fig. 5, we show the results for the three values of J_2/J_1 considered in this work. Again, the agreement with DMRG calculations is excellent in all cases, and the ViT state is able to perfectly reproduce the presence of dimer order.

DeepViT.—The ViT wave function can be systematically improved by stacking multiple transformer layers, i.e., $n_l > 1$. Since the optimization of complex-valued deep networks is difficult with standard protocols, we develop a procedure based on the physical interpretation of the attention weights. We start by setting for each head and layer $\alpha_{i-j} = 0$ if $|i - j| > \text{cut}$, with $\text{cut} < L/b$, training only the remaining weights. Small cut values (e.g., cut $= 1$) are good starting points for stable optimizations. Then the cut is relaxed until reaching L/b, where all-to-all connections among the inputs of each layer are restored. As an example, the results for the Heisenberg model with $L = 40$ are shown in Fig. 6. Here, we take $n_l = 4$ (each layer has $h = 2$ and $d/h = 2$) and perform the optimization stages with cut $= 1, \ldots, 10$. Every time, when the cut is relaxed, the accuracy of the energy improves. We stress that the optimization is performed without Marshall sign prior.

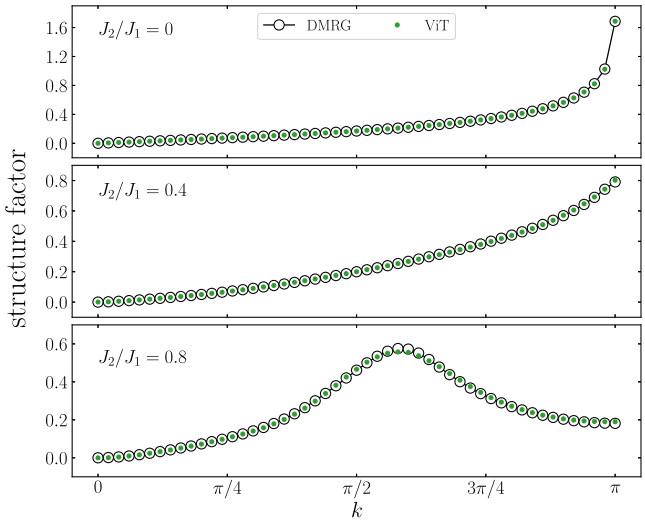

FIG. 4. The spin-spin structure factor $S(k)$ as computed by the ViT wave function (full dots) for $J_2/J_1 = 0$ (upper panel), $J_2/J_1 = 0.4$ (middle panel), and $J_2/J_1 = 0.8$ (lower panel) on a cluster with $L = 100$ sites. The DMRG results are also shown for comparison (empty circles).

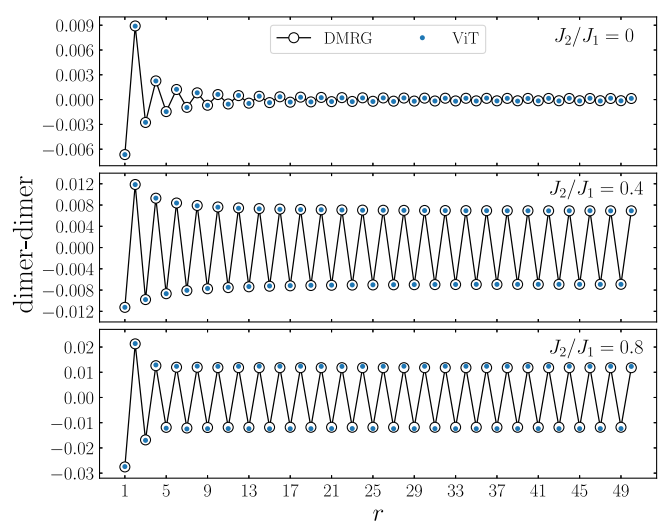

FIG. 5. Dimer-dimer correlations as computed by the ViT wave function (full circles) for $J_2/J_1 = 0$ (upper panel), $J_2/J_1 = 0.4$ (middle panel), and $J_2/J_1 = 0.8$ (lower panel) on a cluster with $L = 100$ sites. The DMRG results are also shown for comparison (empty circles).

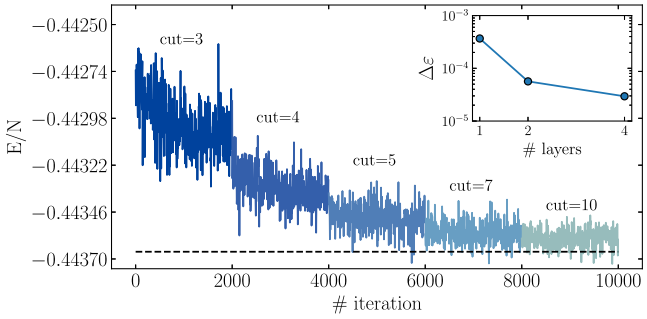

FIG. 6. Optimization of the DeepViT with $n_l = 4$, where each layer has $h = 2$ and $d/h = 2$, for the Heisenberg model with $L = 40$ sites. Along the process, the cut in the attention is fixed and progressively increased from 1 to 10 (the first two values are not shown for better readability). At the end, once the cut has been completely relaxed, the full translational invariance is restored [see Eq. (3)] to compute the accuracy in the energy. Inset: relative error $\Delta\varepsilon$ of the DeepViT wave function by varying the number of layers. The reference energy is computed by DMRG [33] with a bond dimension up to $\chi = 600$ obtaining $E/J_1 = -0.443663$.

Conclusions.—We have introduced a promising class of variational wave functions, which are based upon transformer neural-network architectures (in particular, vision transformers). Their main advantages, with respect to previously defined *Ansätze*, is the mixing of *local* and *global* structures, which makes them very flexible to describe a variety of different quantum phases, with both gapped and gapless spectra. Remarkably, even working with a relatively simple architecture, with $n_l = 1$, excellent results are obtained for a frustrated spin model in one spatial dimension. Generalizations to one-dimensional models with long-range interactions (e.g., the Haldane-Shastry model [39,40]) or two-dimensional models, where ground-state properties are still under debate, are desirable and represent the topic for future investigations, including the calculation of long-range entanglement properties [41]. We expect that for these systems the depth of the network could be important to achieve competitive results with respect to state-of-art numerical methods.

We thank A. Laio and S. Goldt for having drawn our attention to transformers and E. Tirrito for useful discussions about DMRG implementations, which have been performed within the iTensor library [42]. The variational quantum Monte Carlo and the ViT architecture were implemented in JAX [43].

Luciano Loris Viteritti and Riccardo Rende contributed equally to this work.

*Corresponding author.
lucianoloris.viteritti@phd.units.it
†Corresponding author.
rrende@sissa.it

[1] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).
[2] R. B. Laughlin, Phys. Rev. Lett. **50**, 1395 (1983).
[3] G. Carleo and M. Troyer, Science **355**, 602 (2017).
[4] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Phys. Rev. X **8**, 011006 (2018).
[5] K. Choo, T. Neupert, and G. Carleo, Phys. Rev. B **100**, 125124 (2019).
[6] X. Liang, W.-Y. Liu, P.-Z. Lin, G.-C. Guo, Y.-S. Zhang, and L. He, Phys. Rev. B **98**, 104426 (2018).
[7] A. Chen, K. Choo, N. Astrakhantsev, and T. Neupert, Phys. Rev. Res. **4**, L022026 (2022).
[8] C. Roth and A. MacDonald, arXiv:2104.05085.
[9] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, Phys. Rev. Res. **2**, 023358 (2020).
[10] M. Hibat-Allah, R. Melko, and J. Carrasquilla, arXiv: 2207.14314.
[11] D. Luo, Z. Chen, K. Hu, Z. Zhao, V. Hur, and B. Clark, Phys. Rev. Res. **5**, 013216 (2023).
[12] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, Phys. Rev. Lett. **124**, 020503 (2020).
[13] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Phys. Rev. B **96**, 205152 (2017).
[14] F. Ferrari, F. Becca, and J. Carrasquilla, Phys. Rev. B **100**, 125131 (2019).
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[16] 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, arXiv: 2010.11929.
[17] P. Cha, P. Ginsparg, F. Wu, J. Carrasquilla, P. McMahon, and E.-A. Kim, Mach. Learn. **3**, 01LT01 (2021).
[18] D. Luo, Z. Chen, J. Carrasquilla, and B. K. Clark, Phys. Rev. Lett. **128**, 090501 (2022).
[19] S. R. White and I. Affleck, Phys. Rev. B **54**, 9862 (1996).
[20] S. Eggert, Phys. Rev. B **54**, R9612 (1996).
[21] A. Sandvik, AIP Conf. Proc. **1297**, 135 (2010).
[22] S. White, Phys. Rev. Lett. **69**, 2863 (1992).
[23] U. Schollwöck, Ann. Phys. (Amsterdam) **326**, 96 (2011).
[24] F. Becca and S. Sorella, *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University Press, Cambridge, 2017).
[25] W. Marshall, Proc. R. Soc. A **232**, 48 (1955), http://www.jstor.org/stable/99682.
[26] F. Ferrari, A. Parola, S. Sorella, and F. Becca, Phys. Rev. B **97**, 235103 (2018).
[27] L. Viteritti, F. Ferrari, and F. Becca, SciPost Phys. **12**, 166 (2022).
[28] Y. Bengio, P. Simard, and P. Frasconi, IEEE Trans. Neural Networks **5**, 157 (1994).
[29] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, arXiv: 2009.06732.
[30] P. Shaw, J. Uszkoreit, and A. Vaswani, arXiv:1803.02155.
[31] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.130.236401 for details

concerning the optimization of the neural network and the Monte Carlo sampling, which includes Refs. [24,32].
[32] S. Sorella, Phys. Rev. B **71**, 241103(R) (2005).
[33] P. Pippan, S. White, and H. Evertz, Phys. Rev. B **81**, 081103(R) (2010).
[34] T. Westerhout, N. Astrakhantsev, K. Tikhonov, M. Katsnelson, and A. Bagrov, Nat. Commun. **11**, 1593 (2020).
[35] A. Szabó and C. Castelnovo, Phys. Rev. Res. **2**, 033075 (2020).
[36] C.-Y. Park and M. Kastoryano, Phys. Rev. B **106**, 134437 (2022).
[37] M. Bukov, M. Schmitt, and M. Dupont, SciPost Phys. **10**, 147 (2021).
[38] L. Capriotti, F. Becca, A. Parola, and S. Sorella, Phys. Rev. B **67**, 212402 (2003).
[39] F. D. M. Haldane, Phys. Rev. Lett. **60**, 635 (1988).
[40] B. S. Shastry, Phys. Rev. Lett. **60**, 639 (1988).
[41] Y. Zhang, T. Grover, and A. Vishwanath, Phys. Rev. Lett. **107**, 067202 (2011).
[42] M. Fishman, S. White, and M. Stoudenmire, SciPost Phys. Codebases, 4 (2022).
[43] J. Bradbury, R. Frostig, P. Hawkins, M. 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.