

OPEN ACCESS

Are queries and keys always relevant? A case study on transformer wave functions

To cite this article: Riccardo Rende and Luciano Loris Viteritti 2025 *Mach. Learn.: Sci. Technol.* **6** 010501

View the [article online](#) for updates and enhancements.

You may also like

- [A novel dynamic machine learning-based explainable fusion monitoring: application to industrial and chemical processes](#)
Husnain Ali, Rizwan Safdar, Yuanqiang Zhou et al.
- [Asymptotically stable data-driven koopman operator approximation with inputs using total extended DMD](#)
Louis Lortie and James Richard Forbes
- [Large language models for causal hypothesis generation in science](#)
Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou et al.



BENCHMARK

OPEN ACCESS

RECEIVED
8 September 2024REVISED
2 December 2024ACCEPTED FOR PUBLICATION
18 December 2024PUBLISHED
13 January 2025

Original Content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



Are queries and keys always relevant? A case study on transformer wave functions

Riccardo Rende^{1,3,*} and Luciano Loris Viteritti^{2,3} ¹ International School for Advanced Studies, Trieste, Italy² University of Trieste, Trieste, Italy³ Equal contribution.

* Author to whom any correspondence should be addressed.

E-mail: rrende@sissa.it and lucianoloris.viteritti@phd.units.it**Keywords:** neural network quantum states, variational Monte Carlo, vision transformer wave function, attention mechanisms

Abstract

The dot product attention mechanism, originally designed for natural language processing tasks, is a cornerstone of modern Transformers. It adeptly captures semantic relationships between word pairs in sentences by computing a similarity overlap between queries and keys. In this work, we explore the suitability of Transformers, focusing on their attention mechanisms, in the specific domain of the parametrization of variational wave functions to approximate ground states of quantum many-body spin Hamiltonians. Specifically, we perform numerical simulations on the two-dimensional J_1 – J_2 Heisenberg model, a common benchmark in the field of quantum many-body systems on lattice. By comparing the performance of standard attention mechanisms with a simplified version that excludes queries and keys, relying solely on positions, we achieve competitive results while reducing computational cost and parameter usage. Furthermore, through the analysis of the attention maps generated by standard attention mechanisms, we show that the attention weights become effectively input-independent at the end of the optimization. We support the numerical results with analytical calculations, providing physical insights of why queries and keys should be, in principle, omitted from the attention mechanism when studying large systems.

1. Introduction

Transformers [1] have emerged as one of the most powerful deep learning tools in recent years. They are task-agnostic neural networks that are pre-trained to build context-sensitive representations of words in input sentences [2–4]. The success of Transformers lies in their remarkable flexibility: with minimal modifications, they excel in addressing diverse problem domains, often outperforming specialized approaches [5–7]. This is a consequence of their versatile foundational components, namely the dot product self-attention mechanism, Multilayer Perceptron (MLP), Layer Normalization, and skip connections. While elements like the MLP, Layer Normalization, and skip connections are task-agnostic and offer broad applicability, the functional form of the dot product attention mechanism was originally tailored for natural language processing (NLP) tasks. In this context, a sentence is processed by initially associating each word with a vector through a lookup table. These vectors form a sequence which is processed by the self-attention mechanism [1], designed to generate, for each input, an output vector as a weighted sum of all other inputs. Crucially, the coefficients in this sum involve learnable parameters that are optimized to capture the semantic relationships between pairs of words within the sentence. The remarkable generalization properties of Transformers in NLP tasks have been associated with the use of attention weights that depend on the input values, thereby capturing powerful inductive biases related to the semantics in natural languages [8, 9]. One wonders if the dot product attention mechanism provides an inductive bias which is the most appropriate in *any* data domain. For example, [10] in the context of protein contact prediction and [11] in computer vision tasks suggest that input-independent attention weights achieve competitive performance compared to the standard approach. In this paper, we delve into this aspect by exploring the application of the Transformer

architecture as a Neural-Network Quantum State (NQS) for approximating the ground state of quantum many-body spin Hamiltonians on lattice [12]. The Transformer architecture has already been employed in this context, achieving highly accurate results across different systems [13–20]. While many of these works adopt the standard attention mechanism [14, 17, 18], [15] employs a simplified version, omitting queries and keys, still reaching *state-of-the-art* accuracy on one of the most popular benchmark problems in frustrated magnetism. Therefore, the question of whether queries and keys provide a suitable inductive bias for general applications persists. In this work, we tackle this question by systematically investigating the performance of different attention mechanisms within Transformer wave functions. In the following, we summarize our main findings:

- (i) In Transformer wave functions, the standard dot-product attention mechanism used in NLP does not improve the performance of a simpler mechanism in which the attention weights are input-independent.
- (ii) By analyzing the attention maps produced by architectures including queries and keys, we find that the optimization process makes them efficaciously input-independent.
- (iii) Based on analytical computations, we provide insights into why conventional attention mechanisms are expected to converge towards input-independent solutions when applied to systems which are sufficiently large to be split in independent subsystems.

Interestingly, the result of point (iii) can be extended to other domains, such as NLP or computer vision, in cases where tasks can be solved by exploiting correlations over shorter lengths compared to the entire input sequence, thereby partitioning the input into effectively uncorrelated parts.

2. Background

2.1. The quantum-many body problem

The physical properties of an interacting quantum-many body system described by a Hamiltonian \hat{H} are determined by solving the time-independent Schrödinger equation $\hat{H}|\Psi_n\rangle = E_n|\Psi_n\rangle$, where $|\Psi_n\rangle$ and E_n are eigenstates and eigenvalues of \hat{H} , respectively. In principle, fixing a basis in the Hilbert space, we can numerically obtain the spectrum of \hat{H} by storing all its matrix elements and using standard computational routines to diagonalize it. However, a critical challenge arises due to the exponential growth in the size of this matrix with respect to the number of particles in the system, rendering this approach feasible only for small systems [21]. Typically, the focus lies in the low-energy properties of the Hamiltonian, particularly in its ground state $|\Psi_0\rangle$. To obtain approximations of the ground state for systems where exact diagonalization is not feasible, many methods have been developed over the years. Here, we focus on variational approaches, where a variational state $|\Psi_\theta\rangle$, depending on a set of N_p parameters θ , is optimized to minimize the variational energy $E_\theta = \langle\Psi_\theta|\hat{H}|\Psi_\theta\rangle/\langle\Psi_\theta|\Psi_\theta\rangle$. According to the Variational Principle [22], the energy E_θ associated to any generic state $|\Psi_\theta\rangle$ is always bigger than the ground state energy $E_\theta \geq E_0$. Moreover, provided that the ground state is unique, we have that $E_\theta = E_0$ if and only if $|\Psi_\theta\rangle = |\Psi_0\rangle$. To be concrete, we consider systems of N spin-1/2 arranged on regular lattices. In this case, the variational state can be expanded as $|\Psi_\theta\rangle = \sum_{\{\sigma\}} \Psi_\theta(\sigma)|\sigma\rangle$, where $\{\sigma\} = \{\sigma_1^z, \sigma_2^z, \dots, \sigma_N^z\}$ with $\sigma_i^z = \pm 1$ is the computational basis. The many-body wave function $\Psi_\theta(\sigma) = \langle\sigma|\Psi_\theta\rangle$ is a compact representation of the quantum state, which maps configurations of the basis set $|\sigma\rangle$ to complex numbers using a relatively small number of parameters N_p compared to the exponential size of the full Hilbert space ($N_p \ll 2^N$).

2.2. Variational Monte Carlo framework

The Variational Monte Carlo (VMC) is a general framework used to construct an approximation of the ground-state $|\Psi_0\rangle$ of a quantum many-body Hamiltonian \hat{H} [23]. This is achieved by minimizing the variational energy E_θ , associated with a trial variational state $|\Psi_\theta\rangle$, through a gradient-based iterative procedure which employs stochastic estimations of the relevant quantities (see algorithm 1). The key object of the algorithm is the gradient of the energy with respect to the variational parameters (see step 5 in algorithm 1), which can be expressed as a correlation function [12, 15, 23]:

$$F_\gamma = -\frac{\partial E_\theta}{\partial \theta_\gamma} = -2\Re \left[\langle (\hat{H} - \langle \hat{H} \rangle) (\hat{O}_\gamma - \langle \hat{O}_\gamma \rangle) \rangle \right], \quad (1)$$

where $\gamma = 1, \dots, N_p$ and \hat{O}_γ are diagonal operators defined as $O_\gamma(\sigma) = \partial \text{Log}[\Psi_\theta(\sigma)]/\partial \theta_\gamma$. The latter log-derivative can be efficiently computed for NQS architectures using automatic differentiation [24]. The expectation values $\langle \dots \rangle$ are stochastically estimated using Markov Chain Monte Carlo (see appendix A)

Algorithm 1. Variational Monte Carlo.

-
- 1: **Require:** Define a variational state $\Psi_\theta(\sigma)$
 - 2: **Require:** Initialize randomly the variational parameters θ
 - 3: **for** $t = 1, N_{opt}$ **do**
 - 4: samples $\{\sigma_i\}_{i=1}^M \sim |\Psi_\theta(\sigma)|^2$ via MCMC
 - 5: Stochastic estimation of the gradient of the energy: $F_\gamma = -\partial_\gamma E_\theta$ with $\gamma = 1, \dots, N_p$
 - 6: Stochastic estimation of the Quantum Geometric Tensor: $S_{\gamma, \beta}$ with $\gamma, \beta = 1, \dots, N_p$
 - 7: Update of the parameters with Stochastic Reconfiguration: $\delta\theta_\gamma = \tau \sum_\beta S_{\gamma, \beta}^{-1} F_\beta$
 - 8: New parameters: $\theta \leftarrow \theta + \delta\theta$
 - 9: **end for**
-

by sampling M configurations according to the amplitudes $|\Psi_\theta(\sigma)|^2$ (details can be found in appendix B). The parameters are updated according to the Stochastic Reconfiguration (SR) method [25, 26] (see step 7 in algorithm 1), which is formally equivalent to Natural Gradient [27, 28]. The SR approach takes into account the geometric properties of the energy landscape through the Quantum Geometric Tensor S , a $P \times P$ matrix which generalizes the Fisher information metric [29]:

$$S_{\gamma, \beta} = \Re \left[\left\langle \left(\hat{O}_\gamma - \langle \hat{O}_\gamma \rangle \right)^\dagger \left(\hat{O}_\beta - \langle \hat{O}_\beta \rangle \right) \right\rangle \right]. \quad (2)$$

Recent studies have demonstrated the effectiveness of SR in optimizing NQS with a large number of parameters [15, 16, 30]. It is important to stress that in VMC the data, i.e. spin configurations, are generated ‘on the fly’ during the optimization process by sampling from $|\Psi_\theta(\sigma)|^2$. This is different from conventional machine learning scenarios where a fixed training set is provided.

2.3. Vision Transformer wave function

In 2017, Carleo and Troyer [12] proposed using neural networks to parametrize the variational quantum state amplitudes $\Psi_\theta(\sigma) \in \mathbb{C}$. Neural-Network Quantum States have demonstrated remarkable representational power in challenging problems [31, 32] and reached state-of-the-art results in describing the ground state properties of two-dimensional frustrated magnets [15, 16, 30, 33, 34], bosonic [35] and fermionic [36–39] models. In this work, we focus on a particular NQS based on the Vision Transformer (ViT) architecture, introduced in [16]. First, following what is done for images [7], each $L \times L$ input spin configuration σ is split into patches of size $b \times b$, which are linearly embedded in a d -dimensional space, thus producing a sequence of $n = L^2/b^2$ vectors $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, with $\mathbf{x}_i \in \mathbb{R}^d$. This sequence is processed by a deep ViT with real-valued parameters that produce an output sequence of vectors $(\mathbf{y}_1, \dots, \mathbf{y}_n)$, with $\mathbf{y}_i \in \mathbb{R}^d$. The ViT architecture is constituted by n_l encoder blocks, each of them including Multi-Head attention with h heads, two-layer MLP with GeLU activation, skip connections and Pre-Layer Normalization [40]. Then, a d -dimensional hidden representation is obtained as $\mathbf{z} = \sum_{i=1}^n \mathbf{y}_i \in \mathbb{R}^d$. Only at the end, the latter is mapped to a complex number representing the logarithm of the amplitude. This final mapping is performed by an output layer parametrized as a shallow network, namely $\text{Log}[\Psi_\theta(\sigma)] = \sum_{\beta=1}^d g(b_\beta + \mathbf{w}_\beta \cdot \mathbf{z})$, with non-linearity $g(\cdot) = \text{logcosh}(\cdot)$ and complex-valued trainable parameters $\{b_\beta, \mathbf{w}_\beta\}_{\beta=1}^d$. For more details about the architecture see [16].

3. Methods

3.1. Relative positional attention mechanisms

The success of the Transformer architecture is commonly attributed to the attention mechanism [1]. The basic idea of the attention mechanism is to process an input sequence of n vectors $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, with $\mathbf{x}_i \in \mathbb{R}^d$, producing a new sequence $(\mathbf{A}_1, \dots, \mathbf{A}_n)$, with $\mathbf{A}_i \in \mathbb{R}^d$. The goal of this transformation is to construct context-aware output vectors by combining all input vectors [1]:

$$\mathbf{A}_i = \sum_{j=1}^n \alpha_{ij}(\mathbf{x}_i, \mathbf{x}_j) V \mathbf{x}_j. \quad (3)$$

The attention weights $\alpha_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ form a $n \times n$ matrix, where n is the number of patches, which measure the relative importance of the j th input when computing the new representation of the i th input. During the years, several works proposed different parametrizations of the attention weights [44–46]. Here, we consider three different mechanisms, all based on relative positional encoding [44], as appropriate for the objective of this work.

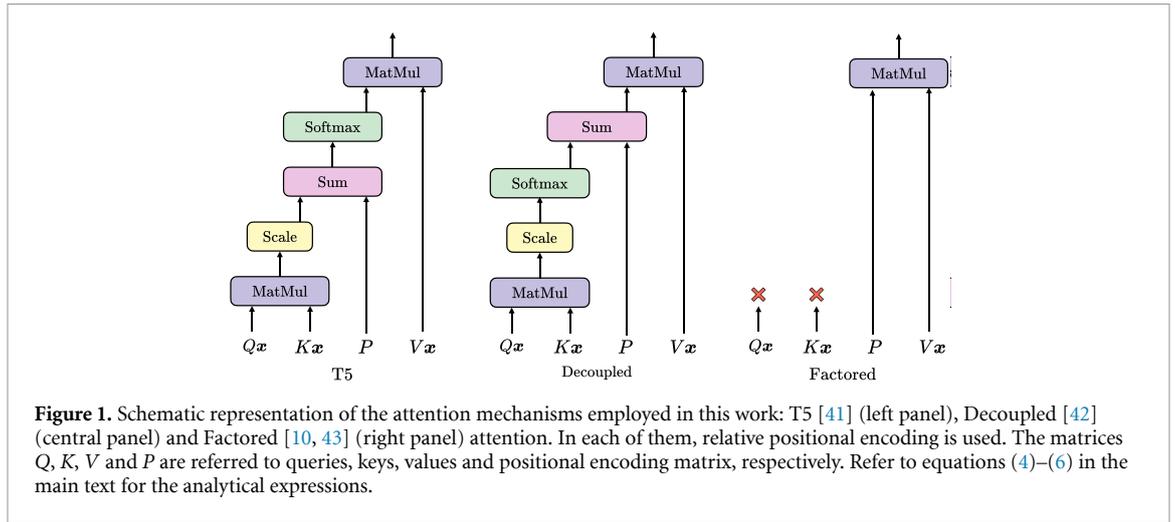


Figure 1. Schematic representation of the attention mechanisms employed in this work: T5 [41] (left panel), Decoupled [42] (central panel) and Factored [10, 43] (right panel) attention. In each of them, relative positional encoding is used. The matrices Q, K, V and P are referred to queries, keys, values and positional encoding matrix, respectively. Refer to equations (4)–(6) in the main text for the analytical expressions.

1. *T5 attention*, introduced in [41], is one of the most popular attention mechanisms:

$$\alpha_{ij}^{T5}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\exp\left(\frac{\mathbf{x}_i^T Q^T K \mathbf{x}_j}{\sqrt{d}} + p_{i-j}\right)}{\sum_{k=1}^n \exp\left(\frac{\mathbf{x}_i^T Q^T K \mathbf{x}_k}{\sqrt{d}} + p_{i-k}\right)}. \quad (4)$$

2. *Decoupled attention*, introduced in [42]:

$$\alpha_{ij}^D(\mathbf{x}_i, \mathbf{x}_j) = \frac{\exp\left(\frac{\mathbf{x}_i^T Q^T K \mathbf{x}_j}{\sqrt{d}}\right)}{\sum_{k=1}^n \exp\left(\frac{\mathbf{x}_i^T Q^T K \mathbf{x}_k}{\sqrt{d}}\right)} + p_{i-j}. \quad (5)$$

3. *Factored attention*, introduced in references [10, 11, 43]:

$$\alpha_{ij}^F(\mathbf{x}_i, \mathbf{x}_j) = p_{i-j}. \quad (6)$$

The vectors $Q\mathbf{x}_i, K\mathbf{x}_i$ and $V\mathbf{x}_i$ are called queries, keys and values, respectively. The matrices Q, K and V , along with the positional encoding P , are trainable parameters. When using relative positional encoding, the matrix P is a *circulant matrix* with dimensions $n \times n$ which is constructed by different circular shifts of a vector of parameters in different rows. This results in only n independent trainable parameters, denoted by p_{i-j} . In figure 1, we show a schematic representation of these three different attention mechanisms. The Factored version has a reduced number of parameters, being the attention weights input independent. Regarding the computational cost for the calculation of each attention weight, we have $O(1)$ complexity in the Factored case and $O(nd^2) + O(n^2d)$ in the other two cases. Decoupled attention, as represented by equation (5), is the simplest extension of the Factored version in equation (6), where the attention weights now factor in the input dependence: setting $Q = K = 0$ allows recovering the Factored attention, albeit with a constant shift. Instead, in T5 attention (see equation (4)) all the attention weights are constrained to be positive due to the global softmax activation.

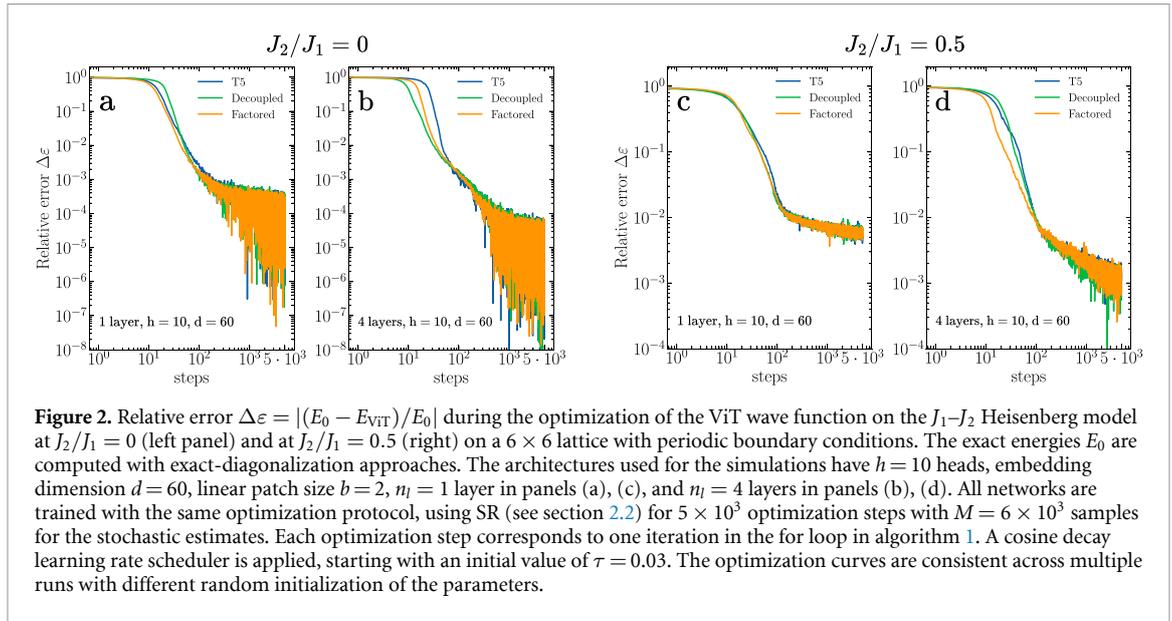
4. Results

4.1. Numerical experiments

We consider the two-dimensional J_1 – J_2 Heisenberg model on a $L \times L$ square lattice, described by the following Hamiltonian:

$$\hat{H} = J_1 \sum_{\langle i,j \rangle} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j, \quad (7)$$

where $\hat{\mathbf{S}}_i = (S_i^x, S_i^y, S_i^z)$ and $J_1, J_2 \geq 0$ are antiferromagnetic couplings for nearest- and next-nearest neighbors, respectively. The ground state of this model exhibits magnetic order in the two distinct limits $J_2/J_1 \ll 1$ and $J_2/J_1 \gg 1$. Specifically, when $J_2 = 0$ ($J_1 = 0$) the model reduces to the unfrustrated Heisenberg model,



characterized by long-range Néel (columnar) magnetic order [47, 48]. In the intermediate region, particularly around $J_2/J_1 \approx 0.5$, the system becomes highly frustrated, giving rise to exotic phases of matter [49]. The determination of the precise nature of the ground state in the frustrated region remains challenging and subject to debate [33, 50, 51].

We employ a ViT wave function (see section 2.3) to approximate, in the VMC framework (see section 2.2), the ground state of this model on a $L \times L$ lattice with periodic boundary conditions. In order to assess the efficacy of the three distinct attention mechanisms introduced in section 3.1, we perform simulations on a 6×6 cluster utilizing ViT architectures with identical hyperparameters (embedding dimension d , number of heads h , number of layers n_l , and linear patch size b), modifying only the attention mechanism, namely T5 (see equation (4)), Decoupled (see equation (5)), and Factored (see equation (6)). In figure 2, we report the optimization curves of the relative error of the variational energy with respect to the exact ground-state energy as a function of the optimization steps. On the left, we present the results for the unfrustrated case ($J_2/J_1 = 0$) using ViT architectures with one (panel (a)) and four (panel (b)) layers. Instead, on the right, we report the results in the frustrated regime ($J_2/J_1 = 0.5$), again using one (panel (c)) and four (panel (d)) layers architectures. We emphasize that, although it is possible to enhance the performance of the variational state by employing larger architectures, such as increasing the number of layers, considering larger embedding dimensions or augmenting the number of heads [15, 16, 19], the use of T5 or Decoupled attention mechanisms with input-dependent attention weights, and the subsequent increase of computational complexity and parameter count via the matrices Q and K , does not produce improved results compared to Factored attention with input-independent attention weights. Notably, not only are the final accuracies practically identical, but also the learning dynamics exhibit similar behavior.

In figure 3, we extend our analysis to larger system sizes, specifically for $L = 8$ and $L = 10$. We focus on an architecture with the same hyperparameters for the different sizes: number of heads ($h = 10$), embedding dimension ($d = 60$), linear patch size ($b = 2$) and number of layers ($n_l = 4$). The left panel displays the relative error of the variational energy as a function of the system size L at $J_2/J_1 = 0.5$. The reference energies used to compute the accuracy are obtained through exact diagonalization for $L = 6$ [52] and through variance extrapolation from [50] and [30], for $L = 8$ and $L = 10$, respectively. This plot demonstrates that the accuracy remains size-consistent across the tested clusters, showing a constant behavior when increasing the system size, despite the fact that the network has fixed complexity. In the right panel, we present the computational time per optimization step as a function of the system size measured on a single GPU A100. The data illustrate how the efficiency gap between the Factored attention mechanism and the other attention mechanisms becomes more pronounced when increasing the system size.

In table 1 we report the results on a 6×6 and a 10×10 lattice at $J_2/J_1 = 0.5$, obtained using a four-layer architecture. In both tables, the first column shows the final mean energy achieved by the different attention mechanisms, the second column indicates the number of parameters employed in the architectures, and the last column presents the total computational time measured on a single GPU A100 to perform 5×10^3 optimization steps. It is worth noting that the accuracy of the results can be further enhanced by restoring

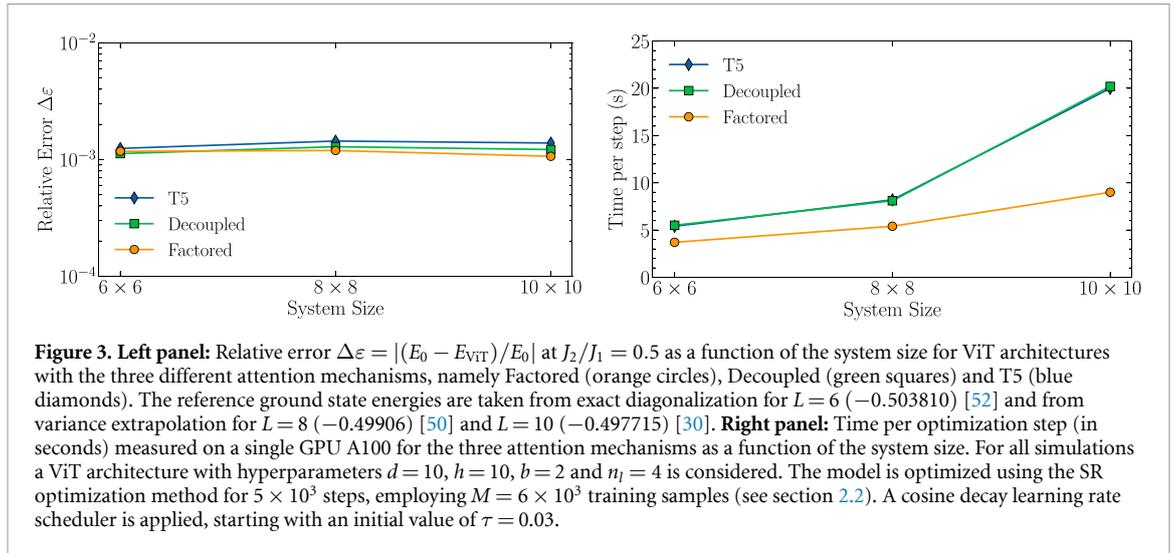


Table 1. Results for the J_2 - J_1 Heisenberg model at $J_2/J_1 = 0.5$ obtained using a ViT architecture with a number of heads $h = 10$, embedding dimension $d = 60$, linear patch size $b = 2$ and a number of layers $n_l = 4$ on a 6×6 (left) and on a 10×10 lattice (right).

	Energy	Parameters	Time		Energy	Parameters	Time
T5	$-0.503182(9)$	184 260	10 h	T5	$-0.497025(6)$	184 900	28 h
Decoupled	$-0.503243(9)$	184 260	10 h	Decoupled	$-0.497108(6)$	184 900	28 h
Factored	$-0.503216(8)$	154 980	6 h	Factored	$-0.497184(6)$	155 620	12.5 h

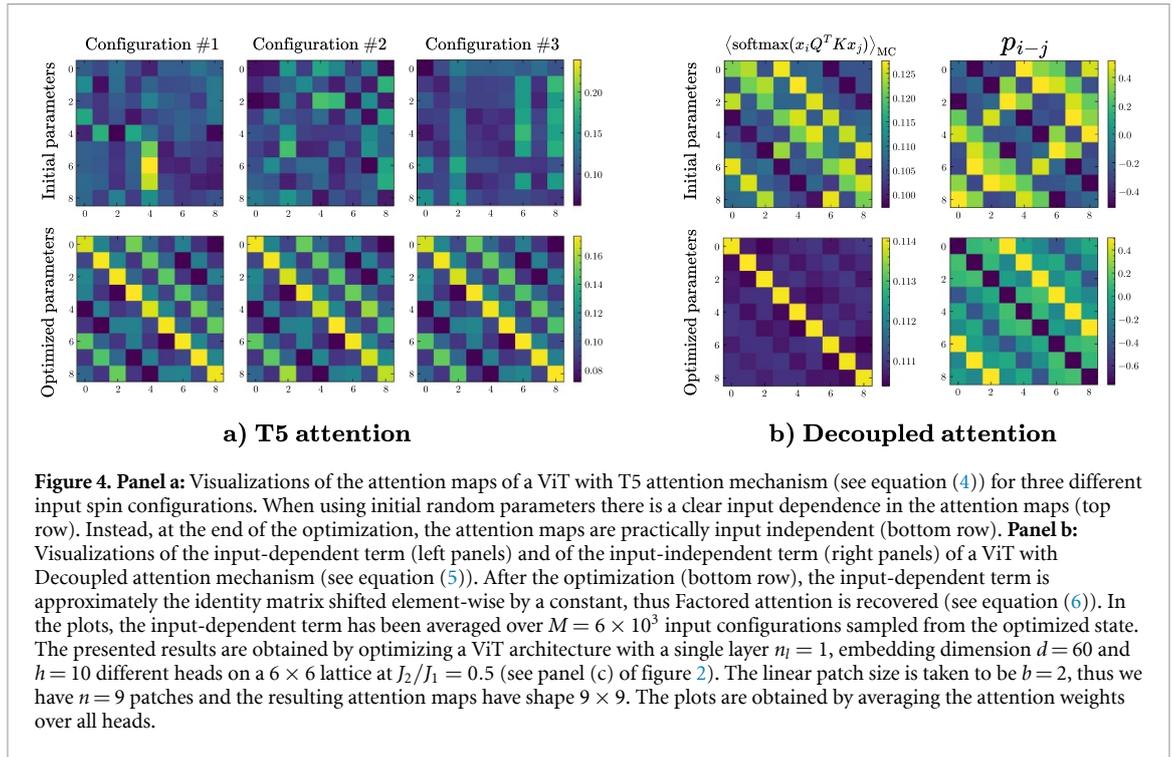
the physical symmetries of the model through quantum number projection approaches [53, 54]; however, this goes beyond the scope of our work.

4.2. Analysis of the attention maps

The main result of the numerical simulations reported in figures 2, 3 and discussed in section 4.1 is that, using a ViT employing T5, Decoupled and Factored attention, the final accuracy is practically the same (see table 1). This suggests that, in the case of T5 and Decoupled attention, queries and keys are effectively not used in the optimized solution. To validate this statement, we study the attention maps. For the analysis, we used a single-layer architecture, where the interpretation of the results is simplified since the patches are only mixed within the attention mechanism, and the subsequent MLP cannot modify the relative weights among the various attention vectors. In panel (a) of figure 4, we consider the case of T5 attention, plotting the attention weights defined in equation (4) for three different input spin configurations. We first check that at the beginning, with random parameters, the attention maps depend on the inputs (top row), ensuring that we have an unbiased initialization. In the bottom row, we show that the architecture after optimization produces input-independent attention maps, thus automatically recovering a positional-only solution. In panel (b) of figure 4, we consider the case of Decoupled attention, plotting separately the input dependent and the positional contributions of the attention weights (see equation (5)). Again, after optimization the network swaps from an unbiased solution (top row) to a positional only solution (bottom row), where the input-dependent term converges approximately to the identity matrix shifted element-wise by a constant. In other words, Factored attention is spontaneously recovered from the Decoupled version (see section 3.1).

4.3. Representation of physical ground states with factored attention

In this section, we provide analytic calculations about the efficacy of input-independent attention mechanisms for approximating quantum states. We first examine an analytically solvable quantum many-body Hamiltonian, developing an exact mapping between its ground state and a single layer of two-headed Factored attention. Building upon this result, we extend our analysis to scenarios where the ground state lacks analytical solutions, providing insights into why attention mechanisms including queries and keys (as in equations (4) and (5)) should converge to positional-only solutions when studying large systems. As an illustrative example of a solvable quantum many-body Hamiltonian, we consider the Shastry-Sutherland model [55], which captures the low-temperature properties of $\text{SrCu}_2(\text{BO}_3)_2$, a compound known for its intriguing physical properties [56]. In a finite range of the frustration ratio, the ground state of this model is represented as a product of singlets between next-nearest-neighbor spins arranged on a square lattice [55], refer to figure 5 for a graphical representation. Here, we want to show that a



single-layer ViT with Factored attention (see equation (6)) can represent exactly this ground state. Working on a $L \times L$ square lattice with periodic boundary conditions, we partition input spin configurations into $b \times b$ patches, with $b = 2$ (see figure 5), which are then flattened to construct input sequences. Assuming an embedding dimension of $d = b^2 = 4$ and choosing the embedding matrix to be the identity, the i th input vector is $\mathbf{x}_i = (\sigma_{i,1}, \sigma_{i,2}, \sigma_{i,3}, \sigma_{i,4})^T$, where $i = 1, \dots, n$, with $n = L^2/b^2$. Then, we apply the Multi-Head attention mechanism [1] with $h = 2$ heads. Considering the value matrices:

$$V^{(1)} = \begin{pmatrix} 0 & 0 & 0 & V_{11}^{(1)} \\ 0 & V_{22}^{(1)} & V_{23}^{(1)} & 0 \end{pmatrix} \quad V^{(2)} = \begin{pmatrix} V_{14}^{(2)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (8)$$

the value vectors are computed as $\mathbf{v}_i^{(\mu)} = V^{(\mu)} \mathbf{x}_i \in \mathbb{R}^{d/h}$:

$$\mathbf{v}_i^{(1)} = \left(V_{11}^{(1)} \sigma_{i,4}, V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3} \right)^T \quad \mathbf{v}_i^{(2)} = \left(V_{14}^{(2)} \sigma_{i,1}, 0 \right)^T. \quad (9)$$

Now, we assume the $n \times n$ attention matrices to be $\alpha_{ij}^{(1)} = \delta_{i,j}$ and $\alpha_{ij}^{(2)} = \delta_{i,S(i)}$, where:

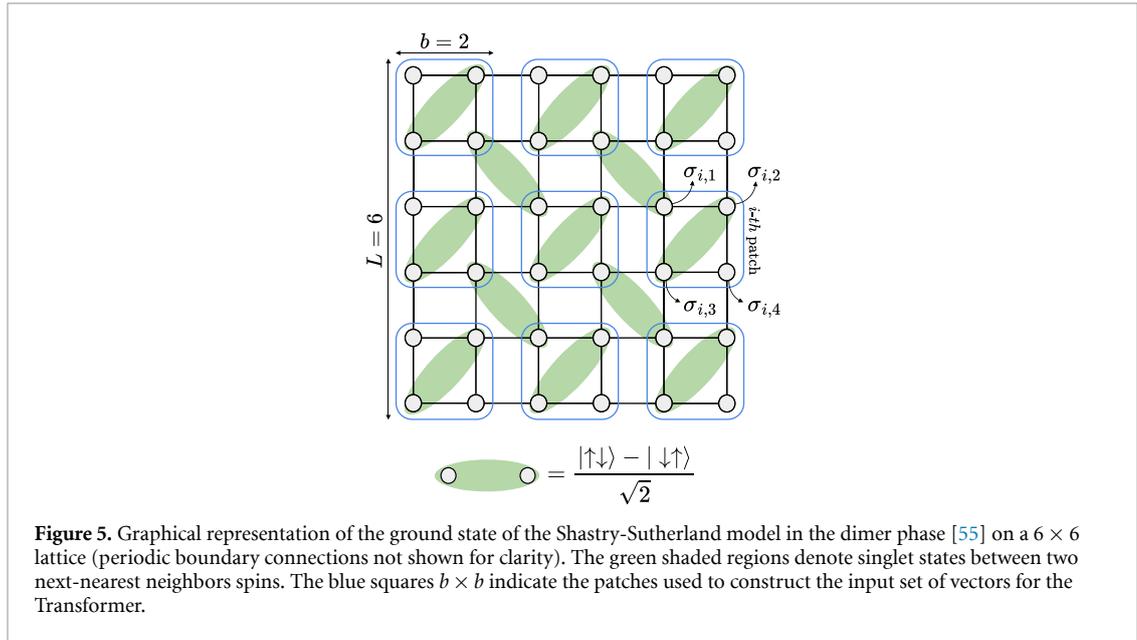
$$S(i) = \begin{cases} (i+1) \% n & \text{if } i \% (L/b) = 0, \\ (i+L/b) \% n + 1 & \text{otherwise,} \end{cases} \quad (10)$$

to take into account the periodic boundary conditions. Notably, the role of the two different heads is to encode the intra-patches correlations through the attention matrix $\alpha^{(1)}$ and the inter-patches correlations through $\alpha^{(2)}$. It is worth noting that, to reproduce the same attention maps with T5 (see equation (4)) or Decoupled (see equation (5)) attention mechanisms, we have to set $Q = K = 0$. The resulting attention vectors are:

$$\mathbf{A}_i^{(1)} = \left(V_{11}^{(1)} \sigma_{i,4}, V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3} \right)^T \quad \mathbf{A}_i^{(2)} = \left(V_{14}^{(2)} \sigma_{S(i),1}, 0 \right)^T. \quad (11)$$

Following the Multi-Head mechanism [1], we concatenate the vectors $\mathbf{A}_i^{(\mu)}$ of the different heads and apply another matrix $W \in \mathbb{R}^{d \times d}$ to mix the different representations. Choosing W to be:

$$W = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (12)$$



we obtain:

$$\mathbf{A}_i = \left(V_{11}^{(1)} \sigma_{i,4} + V_{14}^{(2)} \sigma_{S(i),1}, V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3}, 0, 0 \right)^T. \tag{13}$$

At this point, in the standard architecture each attention vector is fed to a MLP; in our analytical computations, we substitute it with a generic nonlinearity $F(\mathbf{A}_i + c)$, where c is a constant bias. The output of this operation is the sequence of vectors:

$$\mathbf{y}_i = \left(F \left(V_{11}^{(1)} \sigma_{i,4} + V_{14}^{(2)} \sigma_{S(i),1} + c \right), F \left(V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3} + c \right), 0, 0 \right)^T. \tag{14}$$

The hidden representation is obtained by summing all the output vectors $\mathbf{z} = \sum_{i=1}^n \mathbf{y}_i$, where $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{z} = \left(\sum_{i=1}^n F \left(V_{11}^{(1)} \sigma_{i,4} + V_{14}^{(2)} \sigma_{S(i),1} + c \right), \sum_{i=1}^n F \left(V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3} + c \right), 0, 0 \right)^T. \tag{15}$$

Replacing the fully-connected network that acts on \mathbf{z} [15, 16, 57] with a simpler sum, we get the amplitude of the input spin configuration:

$$\text{Log}[\Psi_\theta(\sigma)] = \sum_{i=1}^n \left[F \left(V_{11}^{(1)} \sigma_{i,4} + V_{14}^{(2)} \sigma_{S(i),1} + c \right) + F \left(V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3} + c \right) \right]. \tag{16}$$

At the end, by choosing $F(\cdot) = \text{logcos}(\cdot)$ and setting $V_{11}^{(1)} = V_{23}^{(1)} = \pi/4$, $V_{14}^{(2)} = V_{22}^{(1)} = 3\pi/4$ and $c = \pi/2$ we obtain an exact representation that fully complies with the ground state of the model, specifically a product of singlets arranged on a square lattice, as illustrated in figure 5:

$$\Psi_0(\sigma) = \prod_{i=1}^{L^2/4} \cos \left(\frac{\pi}{2} + \pi (\sigma_{i,4} + 3\sigma_{S(i),1}) \right) \cos \left(\frac{\pi}{2} + \pi (\sigma_{i,2} + 3\sigma_{i,3}) \right). \tag{17}$$

We want to emphasize that, to keep the analytical calculation manageable, we did not to include Layer Norm and skip connections. The mapping between the exact ground state of the Shastry-Sutherland model and the Transformer wave function highlights the role played by the different components of the architecture. In particular, this example reveals that the attention weights are used to describe the correlations in the ground state, and the attention weights connecting two patches containing uncorrelated spins should be zero to have an exact representation of the ground state.

In general, physical events that are sufficiently far apart (either in space or time) are essentially independent or uncorrelated. From a mathematical perspective, this fundamental concept is formalized through the *cluster property* [58, 59]:

$$\lim_{|i-j| \rightarrow +\infty} \langle \hat{B}_i \hat{B}_j \rangle = \langle \hat{B}_i \rangle \langle \hat{B}_j \rangle, \quad (18)$$

where \hat{B}_i is a generic local operator. According to the cluster property, correlations must decay with distance and, in the thermodynamic limit, sites that are infinitely distant become uncorrelated. As shown in the previous mapping, the role of the attention weights is to connect correlated inputs. Therefore, for systems for which the property in equation (18) holds, we expect the attention weights connecting spins far apart in the system to be close to zero, regardless of the specific values of the spins. Interestingly, to reproduce this long-distance behavior using standard T5 (equation (4)) or Decoupled (see equation (5)) attention mechanisms we have to require $Q = K = 0$. In other words, the standard attention mechanisms should converge to positional only solutions, thereby to the Factored version (see equation (6)). This argument, which exploits only the correlations among the elements of the input sequence, can be extended to any domain provided that the input sequence is long enough that correlations decay significantly within the scale of the system. For example, even in NLP or in computer vision tasks, when considering long input sentences or large images, it must be true that words or patches of pixels that are really far apart are uncorrelated, and so in this limit queries and keys should be optimized to zero. However, when dealing with finite sequences, this argument can have a marginal impact, and using input dependent attention weights as in equation (4) can provide a good inductive bias for solving the task.

5. Conclusion

In this work, we showed that, when training a Transformer to approximate ground states of quantum many-body Hamiltonians, the standard attention mechanism yields equivalent performance to a simplified version, the Factored attention. The latter utilizes input-independent attention weights, resulting in fewer parameters and reduced computational cost. Moreover, starting from analytical computations, we established a direct link between attention weights and correlations. We observed that if the dominant correlation lengths necessary to solve a specific task are shorter than the total input size, the weights in conventional attention mechanisms (e.g. T5 [41]) should converge towards input-independent solutions. Interestingly, the same considerations can be extended to NLP and computer vision domains. For example, in image classification tasks, the pertinent scale is associated with the extension of objects requiring detection, typically smaller than the entire image. A straightforward approach to mitigate potential problems associated with the relationship between long-range behavior of correlations and queries and keys is the implementation of *local* attention mechanisms, wherein attention weights beyond a specified distance are manually set to zero. In [60] it has been found that it is possible to use short-range attention for the majority of layers in the Transformer and recover the same performance of long-range language modeling. However, we emphasize that a necessary condition for the validity of our results is the possibility to partition the input sequences into effectively uncorrelated segments. This requirement may not hold universally across NLP applications. For instance, studies have demonstrated that correlations can extend over arbitrarily long scales in literary texts [61, 62], and that, for specific tasks, global token mechanisms are preferred [63]. An interesting future direction of research could be the design of attention mechanisms that are able to describe the decay of long-range correlations without the necessity to set queries and keys to zero or without employing local attention mechanisms.

Reproducibility

The variational quantum Monte Carlo and the ViT architecture were implemented in JAX [24]. The implementation of the Stochastic Reconfiguration [15] is available on NetKet [64] under the name of `VMC_SRT`. The ViT architecture used in this paper is available at: <https://zenodo.org/records/14060431>.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

We thank A Laio and F Becca for useful discussions. We acknowledge the CINECA award under the ISCRa initiative, for the availability of high-performance computing resources and support.

Appendix A. Monte Carlo expectation values

The expectation value of a quantum operator \hat{B} on a variational state $|\Psi_\theta\rangle$ can be computed as

$$\langle \hat{B} \rangle = \frac{\langle \Psi_\theta | \hat{B} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle} = \sum_{\{\sigma\}} P_\theta(\sigma) B_L(\sigma), \quad (19)$$

where $P_\theta(\sigma) = |\Psi_\theta(\sigma)|^2 / \langle \Psi_\theta | \Psi_\theta \rangle$ and $B_L(\sigma) = \langle \sigma | \hat{B} | \Psi_\theta \rangle / \langle \sigma | \Psi_\theta \rangle$ is the so-called *local estimator* of \hat{B} . The previous expression allows us to introduce a controlled approximation method for computing expectation values. Specifically, we can perform a stochastic estimation:

$$\bar{B} = \frac{1}{M} \sum_{i=1}^M B_L(\sigma_i), \quad (20)$$

with $\{\sigma_1, \sigma_2, \dots, \sigma_M\}$ generated from the distribution $P_\theta(\sigma)$ (see appendix B). The accuracy of the estimation is controlled by a statistical error which scales as $O(1/\sqrt{M})$.

It is important to note that the computation of the local estimator $B_L(\sigma)$ in principle requires a summation over an exponential number of terms in the system size:

$$B_L(\sigma) = \sum_{\{\sigma'\}} \langle \sigma | \hat{B} | \sigma' \rangle \frac{\Psi_\theta(\sigma')}{\Psi_\theta(\sigma)}. \quad (21)$$

However, for local operators, such as the Hamiltonian, $B_L(\sigma)$ can be computed efficiently. This is because the number of connected configurations σ' for which $\langle \sigma | \hat{B} | \sigma' \rangle \neq 0$ scales polynomially with the system size.

Appendix B. Metropolis algorithm

The Metropolis algorithm allows the generation of a Markov Chain [23] of configurations $\{\sigma_1, \sigma_2, \dots, \sigma_M\}$ that are distributed according to $P_\theta(\sigma) = |\Psi_\theta(\sigma)|^2 / \langle \Psi_\theta | \Psi_\theta \rangle$, without the knowledge of the normalization constant $\langle \Psi_\theta | \Psi_\theta \rangle$. Let us assume that σ is the current configuration of the Markov chain. To obtain the new configuration according to the Metropolis algorithm, we perform the following steps:

1. Generate a configuration $\sigma' \sim k(\sigma' | \sigma)$, where $k(\sigma' | \sigma)$ is the *proposal kernel* [23].
2. Evaluate the log-acceptance ratio of the proposed move:

$$\log[A(\sigma', \sigma)] = \min\left(0, \log\left[\frac{P_\theta(\sigma')}{P_\theta(\sigma)}\right]\right), \quad (22)$$

where

$$\log\left[\frac{P_\theta(\sigma')}{P_\theta(\sigma)}\right] = 2\Re\{\text{Log}[\Psi_\theta(\sigma')]\} - 2\Re\{\text{Log}[\Psi_\theta(\sigma)]\}. \quad (23)$$

3. Accept the new configuration σ' with probability $A(\sigma', \sigma)$. In practice, this is done by drawing a random number $u \in (0, 1]$ and proceeding as follows:
 - **Accept** the move if $\log(u) \leq \log[A(\sigma', \sigma)]$;
 - **Reject** the move if $\log(u) > \log[A(\sigma', \sigma)]$, in this case the new configuration in the Markov Chain remains σ .

Notice that the described formulation of the Metropolis algorithm relies solely on the logarithm of the wave function $\text{Log}[\Psi_\theta(\sigma)]$. This is useful from a practical standpoint to avoid numerical issues, such as underflow and overflow, when evaluating the non-normalized wave function.

In the case of the J_1 - J_2 Heisenberg model studied in this work, due to the $SU(2)$ spin symmetry of the Hamiltonian, the total magnetization is conserved and the ground-state search can be limited in the $S^z = 0$ sector. This can be implemented in the Monte Carlo sampling by proposing the flipping of two spins oriented in opposite directions when generating the new configuration σ' (see step (1) of the Metropolis algorithm).

ORCID iDs

Riccardo Rende  <https://orcid.org/0000-0001-5656-4241>

Luciano Loris Viteritti  <https://orcid.org/0009-0004-2332-7943>

References

- [1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* vol 30 ed I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (Curran Associates, Inc) (available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
- [2] Devlin J, Chang M-W, Lee K and Toutanova K 2019 Bert: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)
- [3] Radford A, Narasimhan K, Salimans T and Sutskever I 2018 Improving language understanding by generative pre-training
- [4] Radford A, Jeffrey W, Child R, Luan D, Amodei D and Sutskever I 2019 Language models are unsupervised multitask learners *OpenAI blog* 1 9 (available at: <https://openai.com/index/language-unsupervised/>)
- [5] Jumper J *et al* 2021 Highly accurate protein structure prediction with alphafold *Nature* **596** 1–11
- [6] OpenAI 2024 Gpt-4 technical report (arXiv:2303.08774)
- [7] Dosovitskiy A *et al* 2021 An image is worth 16x16 words: transformers for image recognition at scale (arXiv:2010.11929)
- [8] Clark K, Khandelwal U, Levy O and Manning C D 2019 What does bert look at? an analysis of bert's attention (arXiv:1906.04341)
- [9] Rogers A, Kovaleva O and Rumshisky A 2021 A primer in BERTology: what we know about how bert works *Trans. Assoc. Comput. Linguist.* **8** 842–66
- [10] Bhattacharya N, Thomas N, Rao R, Dauparas J, Koo P K, Baker D, Song Y S and Ovchinnikov S Interpreting Potts and Transformer protein models through the lens of simplified attention pp 34–45
- [11] Jelassi S, Sander M and Yuanzhi Li 2022 Vision transformers provably learn spatial structure *Advances in Neural Information Processing Systems* vol 35, eds Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (Curran Associates, Inc.) pp 37822–36
- [12] Carleo G and Troyer M 2017 Solving the quantum many-body problem with artificial neural networks *Science* **355** 602–6
- [13] Melko R G and Carrasquilla J 2024 Language models for quantum simulation *Nat. Comput. Sci.* **4** 11–18
- [14] Sprague K and Czischek S 2023 Variational monte carlo with large patched transformers (arXiv:2306.03921)
- [15] Rende R, Viteritti L L, Bardone L, Becca F and Goldt S 2024 A simple linear algebra identity to optimize large-scale neural network quantum states *Commun. Phys.* **7** 260
- [16] Viteritti L L, Rende R, Parola A, Goldt S and Becca F 2023 Transformer wave function for the shastry-sutherland model: emergence of a spin-liquid phase (arXiv:2311.16889)
- [17] Luo Di, Chen Z, Carrasquilla J and Clark B K 2022 Autoregressive neural network for simulating open quantum systems via a probabilistic formulation *Phys. Rev. Lett.* **128** 090501
- [18] Luo Di, Chen Z, Kaiwen H, Zhao Z, Miyoung Hur V and Clark B K 2023 Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models *Phys. Rev. Res.* **5** 013216
- [19] Viteritti L L, Rende R and Becca F 2023 Transformer variational wave functions for frustrated quantum spin systems *Phys. Rev. Lett.* **130** 236401
- [20] von Glehn I, Spencer J S and Pfau D A self-attention ansatz for *ab-initio* quantum chemistry 2023 (arXiv:2211.13672)
- [21] Sandvik A W 2010 Computational studies of quantum spin systems *AIP Conf. Proc.* **1297** 135–338
- [22] Sakurai J J and Napolitano J 2020 *Modern Quantum Mechanics* 3 edn (Cambridge University Press)
- [23] Becca F and Sorella S 2017 *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University Press)
- [24] Bradbury J *et al* 2018 JAX: composable transformations of Python+NumPy programs
- [25] Sorella S 1998 Green function monte carlo with stochastic reconfiguration *Phys. Rev. Lett.* **80** 4558–61
- [26] Sorella S 2005 Wave function optimization in the variational monte carlo method *Phys. Rev. B* **71** 241103
- [27] Amari S and Douglas S C 1998 Why natural gradient? *Proc. 1998 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No.98CH36181) (vol 2)* pp 1213–6
- [28] Amari S, Karakida R and Oizumi M 2019 Fisher information and natural gradient learning in random deep networks *Proc. 22nd Int. Conf. on Artificial Intelligence and Statistics (Proc. Machine Learning Research, vol 89) (16–18 April 2019)* ed Chaudhuri K and Sugiyama M (PMLR) pp 694–702
- [29] Park C-Y and Kastoryano M J 2020 Geometry of learning neural quantum states *Phys. Rev. Res.* **2** 023232
- [30] Chen A and Heyl M 2023 Efficient optimization of deep neural quantum states toward machine precision *Nat. Phys.* **20** 1476–81 (available at: www.nature.com/articles/s41567-024-02566-1)
- [31] Glasser I, Pancotti N, August M, Rodriguez I D and Ignacio Cirac J 2018 Neural-network quantum states, string-bond states and chiral topological states *Phys. Rev. X* **8** 011006
- [32] Lange H, Van de Walle A, Abedinnia A and Bohrdt A 2024 From architectures to applications: a review of neural quantum states *Quantum Sci. Technol.* **9** 040501
- [33] Nomura Y and Imada M 2021 Dirac-type nodal spin liquid revealed by refined quantum many-body solver using neural-network wave function, correlation ratio and level spectroscopy *Phys. Rev. X* **11** 031034
- [34] Roth C, Szabó A and MacDonald A H 2023 High-accuracy variational monte carlo for frustrated magnets with deep neural networks *Phys. Rev. B* **108** 054410
- [35] Denis Z and Carleo G 2024 Accurate neural quantum states for interacting lattice bosons (arXiv:2404.07869)
- [36] Robledo Moreno J, Carleo G, Georges A and Stokes J 2022 Fermionic wave functions from neural-network constrained hidden states *Proc. Natl Acad. Sci.* **119** e2122059119
- [37] Kim J, Pescia G, Fore B, Nys J, Carleo G, Gandolfi S, Hjorth-Jensen M and Lovato A 2023 Neural-network quantum states for ultra-cold fermi gases (arXiv:2305.08831)
- [38] Pfau D, Spencer J S, Matthews A G D G and Foulkes W M C 2020 *Ab initio* solution of the many-electron schrödinger equation with deep neural networks *Phys. Rev. Res.* **2** 033429
- [39] Nys J, Pescia G and Carleo G 2024 *Ab-initio* variational wave functions for the time-dependent many-electron schrödinger equation (arXiv:2403.07447)

- [40] Xiong R, Yang Y, He Di, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L and Liu T-Y 2020 On layer normalization in the transformer architecture (arXiv:2002.04745)
- [41] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Wei Li and Liu P J 2023 Exploring the limits of transfer learning with a unified text-to-text transformer (arXiv:1910.10683)
- [42] Dai Z, Liu H, Quoc V L and Tan M 2021 Coatnet: Marrying convolution and attention for all data sizes (arXiv:2106.04803)
- [43] Rende R, Gerace F, Laio A and Goldt S 2024 Mapping of attention mechanisms to a generalized potts model *Phys. Rev. Res.* **6** 023057
- [44] Shaw P, Uszkoreit J and Vaswani A 2018 Self-attention with relative position representations (arXiv:1803.02155)
- [45] Wennberg U and Eje Henter G 2021 The case for translation-invariant self-attention in transformer-based language models (arXiv:2106.01950)
- [46] Guolin K, He Di and Liu T-Y 2021 Rethinking positional encoding in language pre-training *Int. Conf. on Learning Representations*
- [47] Calandra Buonaura M and Sorella S 1998 Numerical study of the two-dimensional heisenberg model using a green function monte carlo technique with a fixed number of walkers *Phys. Rev. B* **57** 11446–56
- [48] Anders W S 1997 Finite-size scaling of the ground-state parameters of the two-dimensional heisenberg model *Phys. Rev. B* **56** 11678–90
- [49] Savary L and Balents L 2016 Quantum spin liquids: a review *Rep. Prog. Phys.* **80** 016502
- [50] Wen-Jun H, Becca F, Parola A and Sorella S 2013 Direct evidence for a gapless Z_2 spin liquid by frustrating néel antiferromagnetism *Phys. Rev. B* **88** 060402
- [51] Gong S-S, Zhu W, Sheng D N, Motrunich O I and Fisher. M P A 2014 Plaquette ordered phase and quantum phase diagram in the spin- $\frac{1}{2}J_1 - J_2$ square heisenberg model *Phys. Rev. Lett.* **113** 027201
- [52] Schulz H J, Ziman T A L and Poilblanc D 1996 Magnetic order and disorder in the frustrated quantum heisenberg antiferromagnet in two dimensions *J. Phys. I* **6** 675–703
- [53] Nomura Y 2021 Helping restricted boltzmann machines with quantum-state representation by restoring symmetry *J. Phys.: Condens. Matter* **33** 174003
- [54] Reh M, Schmitt M and Gärtner M 2023 Optimizing design choices for neural quantum states *Phys. Rev. B* **107** 195115
- [55] Shastry B S and Sutherland B 1981 Exact ground state of a quantum mechanical antiferromagnet *Physica B+C* **108** 1069–70
- [56] Zayed M E *et al* 2017 4-spin plaquette singlet state in the shastry-sutherland compound $\text{SrCu}_2(\text{BO}_3)_2$ *Nat. Phys.* **13** 962–6
- [57] Rende R, Goldt S, Becca F and Viteritti L L 2024 Fine-tuning neural network quantum states *Phys. Rev. Research* **6** 043280
- [58] Wichmann E H and Crichton J H 1963 Cluster decomposition properties of the s matrix *Phys. Rev.* **132** 2788–99
- [59] Weinberg S 1999 *What is Quantum Field Theory and What did we Think it was?* (Cambridge University Press) pp 241–51
- [60] Rae J and Razavi A 2020 Do transformers need deep long-range memory? *Proc. 58th Annual Meeting of the Association for Computational Linguistics* ed D Jurafsky, J Chai, N Schluter and J Tetreault (Association for Computational Linguistics) pp 7524–9
- [61] Altmann E G, Cristadoro G and Degli Esposti M 2012 On the origin of long-range correlations in texts *Proc. Natl Acad. Sci.* **109** 11582–7
- [62] Alvarez-Lacalle E, Dorow B, Eckmann J-P and Moses E 2006 Hierarchical structures induce long-range dynamical correlations in written texts *Proc. Natl Acad. Sci.* **103** 7956–61
- [63] Qin G, Feng Y and Van Durme B 2023 The nlp task effectiveness of long-range transformers
- [64] Vicentini F *et al* 2022 NetKet 3: machine learning toolbox for many-body quantum systems *SciPost Phys. Codebases* **7**