SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area - PhD course in
Theoretical and Scientific Data Science

Theory of Transformers and their application to Neural Quantum States

Candidate:
Riccardo Rende

Advisors:
Sebastian Goldt
Alessandro Laio

Academic Year 2024-25



A Luciano

Contents

Abstract **6**

Publications List **13**

1 Mapping of attention mechanisms to a generalized Potts model **16**
1.1 The self-attention layer 16
1.2 Potts and factored attention 18
1.2.1 Generalized Potts model 18
1.2.2 Masked language modeling 19
1.2.3 Training a transformer on the generalized Potts model . . . 19
1.2.4 Factored self-attention learns the generalized Potts model . . 21
1.3 The sample complexity of self-attention 22
1.4 Discussion 25

2 A distributional simplicity bias in the learning dynamics of transformers **26**
2.1 Simplicity biases in Transformers 27
2.2 Expressing many-body distributions with factored attention 29
2.3 Learning many-body interactions with factored attention 31
2.3.1 Analysing the gradient flow for two-layer factored attention . 33
2.4 Sequential learning in NLP 35
2.5 Discussion 40

3 Vision Transformer Wave Function **42**
3.1 Introduction to Neural Quantum States 42
3.2 The ViT wave function 45
3.3 ViT for the J_1-J_2 Heisenberg chain 49
3.4 Relax-cut procedure to optimize deep complex-valued ViT 52
3.5 Discussion 52

4 Feature learning perspective on Neural Quantum States **53**
4.1 Feature learning inspired Ansatz . . . 54
4.1.1 Vision Transformer . . . 55
4.1.2 Encoder Block . . . 57
4.1.3 Hidden representations . . . 58
4.2 A simple linear algebra identity to optimize Large-Scale Neural Network Quantum States . . . 60
4.3 State of the art results on the J_1-J_2 Heisenberg model . . . 62
4.4 Discussion . . . 66

5 The Shastry-Sutherland model **67**
5.1 Emergence of a gapless spin liquid phase in the Shastry-Sutherland model . . . 67
5.2 Numerical results . . . 70
5.2.1 Benchmarks . . . 71
5.2.2 Phase diagram . . . 72
5.2.3 Nature of the Spin Liquid . . . 75
5.2.4 Attention maps . . . 79
5.3 Discussion . . . 79

6 The role of Queries and Keys in Neural Quantum States **80**
6.1 The dot product attention mechanism . . . 80
6.2 Relative positional attention mechanisms . . . 82
6.3 Numerical experiments . . . 83
6.3.1 Analysis of the attention maps . . . 86
6.3.2 Representation of physical ground states with Factored attention . . . 88
6.4 Discussion . . . 91

7 Foundation Neural-Networks Quantum States as a Unified Ansatz for Multiple Hamiltonians **92**
7.1 Foundation models . . . 93
7.2 Theoretical framework for FNQS . . . 95
7.2.1 Stochastic Reconfiguration for multiple systems . . . 96
7.2.2 Foundation Neural-Network architecture . . . 97
7.2.3 Generalized Fidelity Susceptibility . . . 98
7.3 Numerical results . . . 100
7.3.1 Transverse field Ising chain . . . 100
7.3.2 J_1-J_2-J_3 Heisenberg model . . . 102
7.3.3 Random transverse field Ising model . . . 105
7.3.4 Out of distribution generalization . . . 108

7.4 Discussion . . . 110

Future perspectives **110**

Acknowledgements **113**

8 Appendix **115**
8.1 Appendix to Chapter 1 . . . 115
8.1.1 Replica analysis of self-attention . . . 115
8.1.2 Statistical physics formulation of machine learning problems 115
8.1.3 Replica Calculation . . . 116
8.2 Appendix to Chapter 2 . . . 124
8.2.1 Analytical details of the dynamics of a factored attention layer 124
8.2.2 Experimental setup for simplicity bias in Transformers . . . 125
8.3 Appendix to Chapter 3 . . . 128
8.3.1 Variational Monte Carlo . . . 128
8.3.2 Monte Carlo expectation values . . . 128
8.3.3 Metropolis Algorithm . . . 128
8.4 Appendix to Chapter 4 . . . 129
8.4.1 Lattice and symmetries of the Shastry-Sutherland model . . 129
8.4.2 Extrapolations details of the order parameters of the Shastry-Sutherland model . . . 130
8.4.3 ViT study of the Shastry-Sutherland model on large lattice sizes . . . 132
8.4.4 Hidden Representations in Frustrated scenario . . . 133
8.5 Appendix to Chapter 7 . . . 135
8.5.1 FNQS expectation values . . . 135
8.5.2 Derivation of the Stochastic Reconfiguration equation for multiple systems . . . 136
8.5.3 FNQS systematic energy improvement . . . 137
8.5.4 FNQS study of the structure factor for $J_3/J_1 = 0$ . . . 139
8.5.5 Derivation of the generalized fidelity susceptibility . . . 140

Abstract

My PhD research focused on the Transformer architecture, a powerful deep neural network model that has emerged as a cornerstone for solving complex problems in natural language processing, image analysis, signal processing, and beyond. In particular, we studied the learning dynamics of this architecture, and its application to the representation of many-body wavefunctions, the so-called Neural Quantum States. Initially, we investigated the representational capabilities of Transformers by characterizing the statistical structures that a simplified Transformer layer, utilizing the so-called factored attention, is capable of learning. Building on these results, we utilized factored attention in deep Transformers to develop an accurate ansatz for approximating the ground states of quantum many-body Hamiltonians within the variational Monte Carlo framework. In this specific application, factored attention is crucial for achieving accurate results, demonstrating superior performance compared to the standard attention mechanism used in most of the other applications of the Transformers, and in particular in natural language processing. Alongside the development of an efficient optimization method for large-scale neural networks, we achieved state-of-the-art results on the most popular benchmark in Neural Quantum States and addressed complex physical problems that are subjects of ongoing debate. Finally, we developed a framework to train Foundation Neural Quantum States, which are versatile neural network models that approximate quantum wave functions of multiple systems simultaneously, enabling accurate estimates of challenging quantities such as disorder averages and fidelity susceptibility. We envision numerous future directions for this approach, including its extension to quantum dynamics by explicitly modeling time-dependent variational states, as well as its application to the design of novel materials in fermionic systems.

Theory of Transformers

Transformers [1] are neural networks that have revolutionized natural language processing and machine learning, rapidly becoming the most widely used architecture across a broad variety of applications. The Transformer architecture is

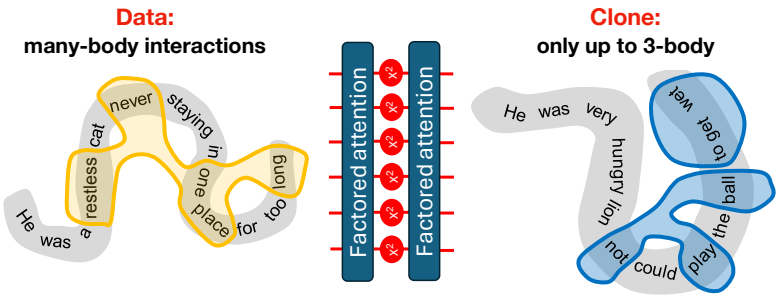


Figure 1: We illustrate the idea of a statistical "clone" of a data set, which is an approximation to the underlying data distribution that keeps only interactions between words up to a fixed degree (in this case, three-body interactions).

built around a fundamental component known as the self-attention mechanism [1], which enables the model to dynamically weigh the relevance of different parts of an input sequence when processing each element. In self-attention, the input sequence of length N is first mapped into a set of representations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, with d being a hyperparameter of the network. Then, each input token is projected into three vectors: a query $\boldsymbol{q}_i = \mathrm{Q}\boldsymbol{x}_i$, a key $\boldsymbol{k}_i = \mathrm{K}\boldsymbol{x}_i$, and a value $\boldsymbol{v}_i = \mathrm{V}\boldsymbol{x}_i$, with Q, K, V matrices of learnable parameters. The output at a given position is computed as a weighted sum of the value vectors from all positions, $\boldsymbol{y}_i = \sum_{j=1}^{N} \alpha(\boldsymbol{q}_i, \boldsymbol{k}_j)\boldsymbol{v}_j$, where the attention weights $\alpha(\boldsymbol{q}_i, \boldsymbol{k}_j)$ are determined by the similarity between the query of the current token and the keys of all tokens in the sequence, and their precise functional form can be chosen according to the task [2]. The output of a self-attention layer is thus a new sequence of the same length N, where each vector $\boldsymbol{y}_i \in \mathbb{R}^d$ integrates long-range dependencies and contextual information from the entire input. By stacking multiple layers of self-attention and feedforward networks, Transformers can learn hierarchical and highly expressive representations of sequences, making them particularly powerful for tasks in natural language processing, computer vision, and beyond.

The key to the success of Transformers is a specific training technique called masked language modeling (MLM). The important feature of this technique is that it can be used with unlabeled data, for example a large amount of text, or several images of various content. Sticking to the example of language, the Transformers are trained to predict missing words in a sentence [3–7]. By learning the conditional distribution of having a word in a specific position of the sentence, given the other words, Transformers ostensibly learn the relationships between words in a robust way. Despite the practical success of Transformers, it remains unclear what type of data distribution self-attention can learn efficiently.

• **Mapping of attention mechanisms to a generalized Potts model.** In Chapter 1, we investigate the statistical structures that self-attention learns with MLM, considering a simplified version called *factored attention*. We are able to show rigorously the equivalence of a factored attention layer trained under masked language modeling and the so-called pseudo-likelihood methods [8–10] for solving the inverse Potts problem. The pseudo-likelihood method is statistically consistent [8–10], i.e. its parameter estimates converge to the true parameters as the number of samples goes to infinity. A direct consequence of this mapping is thus that MLM with factored self-attention enjoys the same asymptotic optimality of pseudo-likelihood methods. Using this mapping, we compute the generalization error of self-attention in a model scenario analytically using the replica method. These findings make factored attention a powerful, theoretically driven building block for deep Transformers, forming the basis for several subsequent developments in this thesis.

• **A distributional simplicity bias in the learning dynamics of Transformers.** Building on the result that training a single layer of factored self-attention is exactly equivalent to solving the inverse Potts problem via the pseudolikelihood method, in Chapter 2 we show that Transformers trained on natural language data display a *simplicity bias*. Simplicity biases have been proposed to explain the success of neural networks, suggesting that neural networks learn increasingly complex functions of their data, going from linear functions to more complex, non-linear functions [11–15]. While simplicity biases have been described theoretically and experimentally in feed-forward networks for supervised learning, the extent to which they also explain the remarkable success of Transformers trained with self-supervised techniques was not studied in depth before our contribution. In Chapter 2 we show that Transformers trained on MLM *sequentially learn many-body interactions* among input words, going from lower to higher orders as training progresses. In this context, factored attention is used to generate clones of a real natural language dataset with a truncated degree of interaction between words (see Fig. 1.1). A rigorous control over the maximum degree of interaction is achieved by controlling the number of layers of a specific architecture employing factored self-attention. This approach opens up the possibilities of studying how interactions of different orders in the data affect learning, in natural language processing and beyond.

Transformer wave function

Building on the theoretical foundation developed in the first two chapters, we introduce a powerful variational ansatz based on Transformer architectures for ap-

proximating ground states of quantum many-body Hamiltonians. The low-energy properties of such systems are governed by the eigenstates of the Hamiltonian operator, with particular interest typically placed on its ground state. While exact diagonalization methods can, in principle, yield the full spectrum of a quantum system by explicitly constructing and diagonalizing the Hamiltonian matrix, this approach becomes impractical for large systems due to the exponential scaling of the Hilbert space with the number of particles. To address this challenge, a variety of approximate methods have been proposed over the years. This thesis focuses on variational approaches, where a trial quantum state, parameterized by a finite set of variables, is optimized to minimize the energy expectation value. According to the variational principle, this energy provides an upper bound to the true ground-state energy, and equality is achieved only when the trial state coincides with the exact ground state. Throughout this thesis, we consider systems composed of spin-1/2 particles arranged on regular lattices. In this setting, the variational state is expressed as a complex-valued wave function defined over classical spin configurations. The key idea is to represent this wave function using a compact neural architecture, such as a Transformer, that can efficiently capture the structure of the quantum state using a number of parameters that scales polynomially, rather than exponentially, with the system size. To optimize these variational parameters, we employ the Variational Monte Carlo (VMC) method [16]. This involves iteratively updating the parameters through gradient-based techniques, where the gradients of the variational energy are estimated stochastically via Monte Carlo sampling. The resulting framework enables the accurate approximation of ground-state properties for quantum systems that are otherwise intractable using exact methods.

In 2017, Carleo and Troyer [17] introduced the idea of using neural networks to parametrize the amplitudes of variational quantum states. Since then, Neural-Network Quantum States (NQS) have demonstrated remarkable expressive power across a wide range of challenging problems [18, 19], achieving state-of-the-art results in the description of ground-state properties for two-dimensional frustrated magnets [20–23], as well as for bosonic [24] and fermionic [25–28] systems. Beginning with Restricted Boltzmann Machines (RBMs)[17]—the simplest NQS architecture, consisting of a single fully connected hidden layer—subsequent research explored a variety of neural architectures, including Convolutional Neural Networks (CNNs)[29–32], Recurrent Neural Networks (RNNs)[33, 34], and Autoregressive Networks[35, 36]. Despite the widespread success of the Transformer architecture in fields such as natural language processing and computer vision, it had not been applied to the parametrization of many-body quantum wave functions. In this thesis, we fill this gap by introducing the Vision Transformer (ViT) wavefunction, demonstrating that the Transformer architecture can effectively represent the ex-

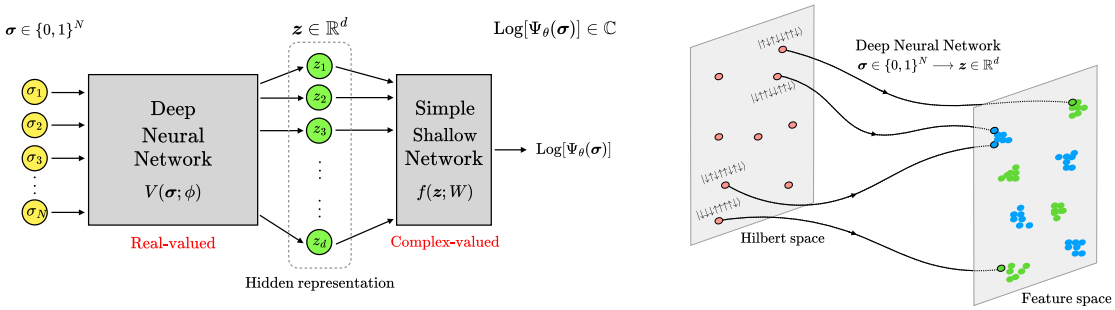


Figure 2: Left panel: The neural quantum state is defined as the composition of two functions: first, a deep neural network $V(\boldsymbol{\sigma};\phi)$ (with real-valued parameters) maps the input configurations $\boldsymbol{\sigma}$ into hidden representations $\boldsymbol{z}$; then, a simple shallow network $f(\boldsymbol{z};W)$ (with complex-valued parameters) generates the logarithm of the amplitudes $\mathrm{Log}[\Psi_\theta(\boldsymbol{\sigma})]$ starting from hidden representations. Right panel: Pictorial illustration of the mapping process carried out by the deep neural network. During this process, the spin configurations of the Hilbert space $\boldsymbol{\sigma}$ are embedded into a feature space $\boldsymbol{z} \in \mathbb{R}^d$, where the determination of the ground-state properties is simplified.

ponential complexity of many-body wave functions by efficiently capturing the non-trivial correlations of quantum states. Using this architecture, we achieved state-of-the-art performance on standard benchmark problems in the field and explored challenging systems whose physical properties are still under active investigation. A central ingredient to the success of this parameterization is the use of the "factored" attention mechanism that we developed in Chapter 1.

• **Transformer variational wave functions for frustrated quantum spin systems.** In Chapter 3, we propose an adaptation of the ViT architecture to define a new class of variational neural-network states for quantum many-body systems, the ViT wave function. Following the most used approach in the literature of Neural Quantum States [19], we employ complex-valued parameters in the network. In this specific application, the factored attention mechanism is crucial for achieving accurate results, demonstrating superior performance compared to the standard attention mechanism used in natural language processing while having fewer parameters and being more computationally efficient (a detailed comparison will be presented in Chapter 6). Using a shallow architecture with a *single layer of factored self-attention*, excellent accuracies compared to Density Matrix Renormalization Group are obtained on one-dimensional systems.

• **Feature Learning perspectives on Neural Quantum States.** In Chapter 4, we address a key limitation of the ViT wavefunction with complex parameters: its difficulty in scaling to deep architectures, which are essential for achieving high accuracy in two-dimensional quantum systems. To overcome this, we introduce a novel perspective on Neural Quantum States based on feature learning. In this approach, a deep neural network with real-valued parameters, implemented using a Transformer architecture, maps physical spin configurations into a high-dimensional feature space. Within this abstract space, determining ground-state properties becomes significantly simpler, requiring only a final output layer with complex-valued parameters to accurately predict the ground state amplitudes (see Figure 4.1). To the best of our knowledge, we are the first to propose this representation-learning-based view of Neural Quantum States, which enables highly stable optimization of deep architectures. Crucially, the proposed ViT *Ansatz* contains a large number of parameters due to its deep architecture, which makes standard optimization methods impractical. To address this problem, we realized that a simple linear algebra result [37] can be exploited to apply Stochastic Reconfiguration for the optimization of large-scale, deep neural quantum states. Using a deep Transformer, we were able to reach the *state-of-the-art energy on the two dimensional J1-J2 Heisenberg model*, a challenging benchmark in highly-frustrated magnetism. This is a pivotal achievement, as we have achieved superior performance compared to all previous approaches using a wave function based purely on neural networks, which is entirely agnostic to the underlying system's physics. This surpasses even those methods where substantial efforts were made to embed as much physical knowledge of the system as possible into the parameterization of the wave function.

• **Transformer wave function for the Shastry-Sutherland model: emergence of a spin-liquid phase.** In Chapter 5, combining the Transformer wave function and the new formulation of the stochastic reconfiguration method, we address a debated physical problem, namely the presence of the spin liquid phase in the Shastry-Sutherland model, designed to capture the low-temperature behavior of SrCu2(BO3)2 with its intriguing properties [38–40]. We supply strong evidence for the *stabilization of a spin-liquid* between the plaquette and antiferromagnetic phases, showing with simulations up to 18×18 lattices that there exists a small, but finite, region in the phase diagram in which both the antiferromagnetic and plaquette order parameters vanish in the thermodynamic limit. In addition, a direct calculation of the triplet excitation at the Γ point provides compelling evidence for a gapless spin liquid. Remarkably, the characterization of the spin liquid phase within the Shastry-Sutherland model has not been explored in previous studies. Our results, we believe, are interesting because they show that the magnetically

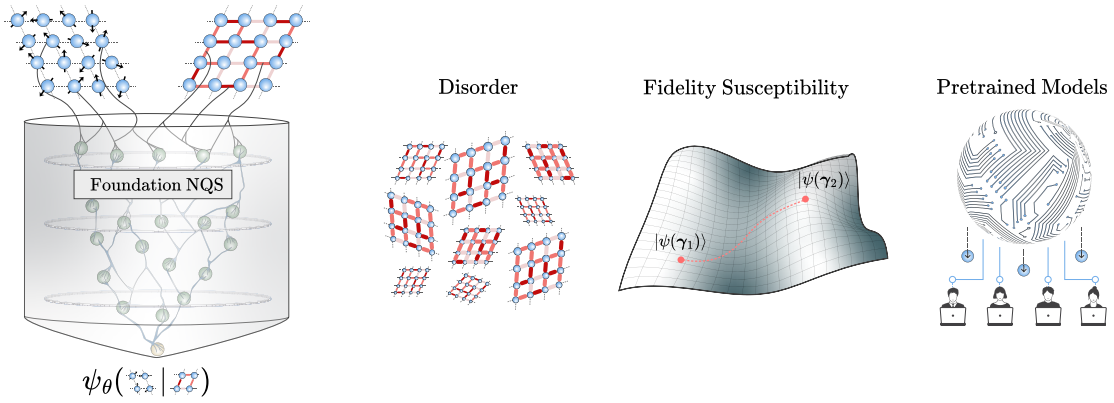


Figure 3: The first panel from the left shows a pictorial representation of Foundation Neural-Network Quantum States (FNQS), which, unlike traditional NQS, process multimodal inputs by incorporating both physical configurations and Hamiltonian couplings to define a variational wave function amplitude over their joint space. FNQS enable a range of applications, including the efficient simulation of disordered systems (see second panel) and the estimation of the quantum geometric tensor in coupling space, also known as the fidelity susceptibility, for the unsupervised detection of quantum phase transitions (see third panel). Moreover, FNQS combined with the public availability of the architectures allows users to leverage pretrained models to explore coupling regimes beyond those encountered during training (see forth panel).

ordered Néel phase is melted into a gapless spin liquid, similarly to what happens in the J_1-J_2 Heisenberg model on the square lattice [41]. This suggests that this kind of (continuous) transition is rather generic and may represent the habit, and not the exception, for the melting of the Néel order due to magnetic frustration. In addition, our findings underscore the potential of the ViT wave function as a valuable tool for probing uncharted phases of matter, and open up new possibilities for establishing the properties of many-body systems.

• **Foundation Neural-Network Quantum States** In Chapter 7, we introduce Foundation Neural-Network Quantum States (FNQS) [42], a new class of variational wave functions that leverage key principles of foundation models. FNQS are built on a single, flexible Transformer architecture capable of processing multimodal inputs, including spin configurations and Hamiltonian physical couplings. In contrast to specialized architectures tailored for specific Hamiltonians, FNQS exhibit strong generalization capabilities, allowing them to accurately represent

quantum states of previously unseen systems. This versatility provides a unified framework adaptable to a broad range of quantum systems and tasks. Beyond ground-state energy estimation, FNQS facilitate the efficient computation of physical quantities that are traditionally challenging or computationally intensive, such as disorder-averaged observables. Furthermore, the fidelity susceptibility can be easily obtained to uncover quantum phase transitions without prior knowledge of order parameters [43]. A visual overview of the FNQS architecture and the range of its applications is presented in Fig. 3.

Publications List

This PhD thesis is based on the following publications.

1. "*Mapping of attention mechanisms to a generalized Potts model*"
   R. Rende, F. Gerace, A. Laio, S. Goldt, Phys. Rev. Research 6, 023057 (2024)

2. "*A distributional simplicity bias in the learning dynamics of transformers*"
   R. Rende, F. Gerace, A. Laio, S. Goldt, NeurIPS 2024

3. "*Transformer variational wave functions for frustrated quantum spin systems*"
   L.L. Viteritti, **R. Rende**, F. Becca, Phys. Rev. Lett. 130, 236401 (2023)

4. "*Are queries and keys always relevant? A case study on Transformer wave functions*"
   R. Rende, L.L. Viteritti, Mach. Learn.: Sci. Technol. 6 (2025) 010501

5. "*Transformer wave function for two dimensional frustrated magnets: Emergence of a spin-liquid phase in the Shastry-Sutherland model*"
   L.L. Viteritti, **R. Rende**, A. Parola, S. Goldt, F. Becca Physical Review B 111 (13), 134411 (2025)

6. "*A simple linear algebra identity to optimize Large-Scale Neural Network Quantum States*"
   R. Rende, L.L. Viteritti, L. Bardone, F. Becca, S. Goldt, Nature Communications Physics 7 (1), 260 (2024)

7. "*Foundation neural-networks quantum states as a unified Ansatz for multiple hamiltonians*"
   R. Rende, L.L. Viteritti, F. Becca, A. Scardicchio, A. Laio, G. Carleo, Nature Communication 16, 7213 (2025)

In addition, the following publications, not included in this thesis, are also part of my PhD work.

1. "*Quantum Spin Glass in the Two-Dimensional Disordered Heisenberg Model via Foundation Neural-Network Quantum States* "
   L.L. Viteritti, **R. Rende**, G. Bracci Testasecca, J. Niedda, R. Moessner, G. Carleo, A. Scardicchio, arXiv:2507.05073 (2025)

2. "*Fine-tuning Neural Network Quantum States*"
   R. Rende, S. Goldt, F. Becca, L.L. Viteritti Phys. Rev. Research 6 (4), 043280 (2024)

3. "*Policy-guided Monte Carlo on general state spaces: Application to glass-forming mixtures*"
   L. Galliano, **R. Rende**, D. Coslovich, The Journal of Chemical Physics 161 (6) (2024)

4. "*An approach to identify the most semantically informative deep representations of text and images*"
   S. Acevedo, A. Mascaretti, **R. Rende**, M. Mahaut, M. Baroni, A. Laio, arXiv:2505.17101 (2025)

Chapter 1

Mapping of attention mechanisms to a generalized Potts model

Transformers are neural networks that revolutionized natural language processing and machine learning. They process sequences of inputs, like words, using a mechanism called self-attention, which is trained via masked language modeling (MLM). In MLM, a word is randomly masked in an input sequence, and the network is trained to predict the missing word. Despite the practical success of transformers, it remains unclear what type of data distribution self-attention can learn efficiently. In this chapter, we show analytically that if one decouples the treatment of word positions and embeddings, a single layer of self-attention learns the conditionals of a generalized Potts model with interactions between sites and Potts colors. More precisely, in Section 1.2 we show that training factored self-attention on the MLM objective is equivalent to solving the inverse Potts problem using the pseudo likelihood method [44–48], and therefore it yields consistent estimators of the parameters. These findings make factored attention a powerful, theoretically driven building block for deep transformers. In Section 1.3, we perform a replica analysis of self-attention, which enables us to compute the generalization loss of the model exactly and yields a non-trivial generalization behavior.

1.1 The self-attention layer

Transformers [49] are a powerful type of neural network that have achieved state-of-the art results in natural language processing (NLP) [3–7], image classification [50], and even protein structure prediction [51]. While standard neural networks can be thought of as functions of a single input, transformers act on sets of "tokens", like words in a sentence. The key to the success of transformers is a training technique called masked language modeling (MLM), where transformers are trained

to predict missing words in a sentence [3–7], see Fig. 1.1. This technique has the advantage that it can leverage large amounts of raw text (or images, or protein sequences) without any annotation. By learning the conditional distribution of having a word in a specific position of the sentence, given the other words, transformers ostensibly learn the relationships between words in a robust way.

The basic building block of transformers is the *self-attention* (SA) mechanism [52, 53], which transforms a sequence of tokens $\mathbf{x}_j$ into another sequence $\mathbf{h}_j$. We illustrate self-attention on a MLM task in Fig. 1.1. The sentence is first transformed into a set of representations $\mathbf{x}_j = \mathbf{e}_j + \mathbf{p}_j$, where $\mathbf{e}_j$ is a vector representing the jth word and the vector $\mathbf{p}_j$ encodes its position. SA then computes a linear transformation of the representations to yield the values $\mathbf{v}_j$. The kth output vector $\mathbf{h}_k$ is then a linear combination of the values $\mathbf{v}_j$ weighted by an attention matrix A, whose elements A_{kj} quantify the relative importance of the jth input token for the kth output vector, for example based on their semantic similarity. The functions to compute values $\mathbf{v}_j$ and the attention matrix A both have trainable parameters; see Eq. (1.2) for a precise definition. The flexibility of self-attention comes from the attention weights A_{kj}, which are not fixed, but computed given the context, i.e. the surrounding tokens.

The practical success of transformers raises several fundamental questions: what are the statistical structures that self-attention learns with MLM? More precisely, since the MLM objective is to learn the conditional probability distribution of words given a set of surrounding words, which family of conditional probabilities can self-attention learn? And how many samples are required to achieve good performance? In this chapter, we make a step towards answering these questions by exploiting tools from the statistical physics of learning [54–57].

The first challenge is to design a data model that mimics the structure of real sentences. While classical works modelled inputs as vectors of i.i.d. random variables, recent work has introduced more sophisticated data models for neural networks [58–64] which allowed the study of unsupervised learning [65, 66]. To analyze the self-supervised learning of MLM, we model sequences of words as a system of spins, interacting via a generalized Potts Hamiltonian [67, 68] with couplings between colors (=words) and positions. We sample a synthetic data set from the Potts model using Monte Carlo, and we perform masked language modeling by training a transformer to predict masked spins in spin sequences. While an off-the-shelf transformer requires several layers of self-attention to learn this simple probability distribution, we show analytically that a single layer of *factored* self-attention, where we separate the treatment of positions and inputs, can reconstruct the couplings of the Potts model *exactly* in the limit of a large training set. In particular, we derive an exact mapping between the output of the self-attention mechanism and the conditional distribution of a Potts spin given the

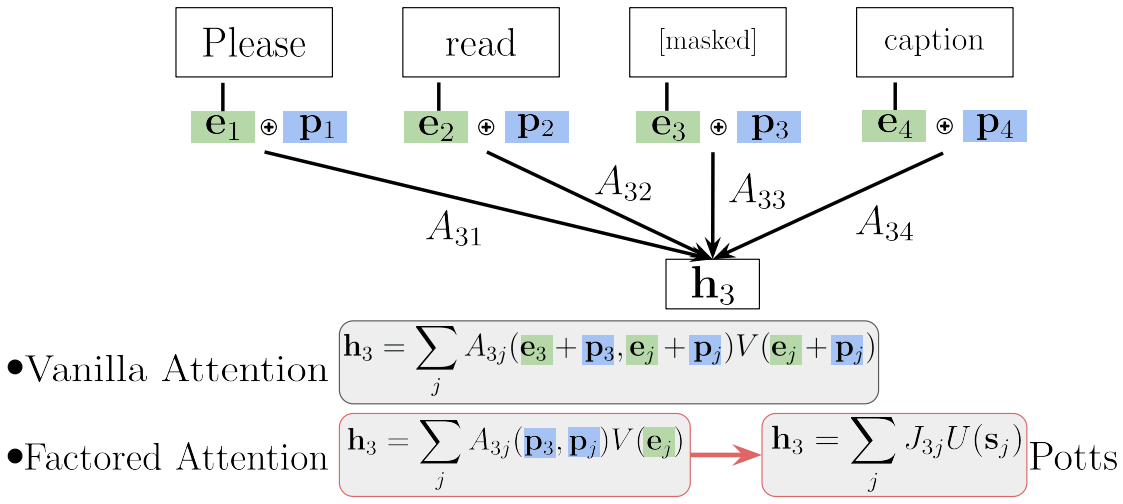


Figure 1.1: **Masked language modeling (MLM) with a single layer of self-attention.** The goal of MLM is to predict the masked word in a given sentence. Self-attention first maps words into representations $\mathbf{e}_j + \mathbf{p}_j$, where $\mathbf{e}_j$ are embedding vectors representing words, and $\mathbf{p}_j$ encode their positions. For a given masked word, the associated attention vector $\mathbf{h}_k$ is computed as a linear combination of the values $\mathbf{v}_j = V(\mathbf{e}_j + \mathbf{p}_j)$ of all other tokens, weighted by the attention weights A_{kj}. In vanilla self-attention, values and attention weights depend on embeddings *and* positional vectors, while in *factored* attention, attention weights depend only on positions, and values only on the embeddings. By identifying the attention weights A with the interaction matrix J of a Potts model Eq. (1.1), the value matrix V with the color similarity matrix U and the embedding vectors with the one-hot spins, we get a learning model identical to a Potts model.

others. We finally use this mapping to compute the generalization loss of a single layer of self-attention analytically using the replica method.

1.2 Potts and factored attention

1.2.1 Generalized Potts model

We model sentences as sequences of spins $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_L)$, with $\mathbf{s}_i \in \mathbb{R}^C$ taking values from a vocabulary of C colors, which we encode as one-hot vectors. Each color can be thought of as a word in natural text, an amino acid in a protein, etc. In a standard Potts Hamiltonian, only spins of the same color interact with each other via an interaction matrix J. This is an unrealistic model for real data: it treats all colors as orthogonal, even though words and amino acids have varying

degrees of similarity. We therefore generalize the Potts Hamiltonian to

$$\mathcal{H}(\mathbf{s}) = -\frac{1}{2}\sum_{i,j=1}^{L} J_{ij}\mathbf{s}_i^T U \mathbf{s}_j, \tag{1.1}$$

where $J \in \mathbb{R}^{L\times L}$ governs the interactions between spins at different positions, and $U \in \mathbb{R}^{C\times C}$ encodes the similarities between colors (we denote matrices by capital letters and vectors in boldface). Without loss of generality, we set $J_{ii} = 0$ and sample sequences from the Boltzmann distribution $P(\mathbf{s}) \propto \exp[-\beta\mathcal{H}(\mathbf{s})]$. We recover the standard Potts model by choosing U as the identity matrix.

1.2.2 Masked language modeling

Given the generative model (1.1), the MLM objective amounts to predicting the ith spin given the sequence $\mathbf{s}_{\setminus i}$ where that spin is "masked", i.e. $\mathbf{s}_i = \mathbf{t}$, the masking token. To apply self-attention to a sequence $\mathbf{s}_{\setminus i}$, we first compute the values $\mathbf{v}_j = V(E\mathbf{s}_j + a\mathbf{p}_j)$, where the embedding matrix $E \in \mathbb{R}^{d\times C}$ maps Potts colors into d-dimensional representation vectors, and $V \in \mathbb{R}^{d\times d}$ is a weight matrix; both E and V are trainable parameters. The scalar parameter a controls the relative importance between the embedding and positional encoding vectors. The output vector $\mathbf{h}_i$ corresponding to the masked token is a linear combination of the values, weighted by an exponential attention function [49]:

$$\mathbf{h}_i\left(\mathbf{s}_{\setminus i}\right) = \sum_{j=1}^{L} \frac{\exp\left[\left(E\mathbf{t}+\mathbf{p}_i\right)^\top Q^\top K\left(aE\mathbf{s}_j+\mathbf{p}_j\right)\right]}{\sum_k \exp\left[\left(E\mathbf{t}+\mathbf{p}_i\right)^\top Q^\top K\left(aE\mathbf{s}_k+\mathbf{p}_k\right)\right]}\mathbf{v}_j. \tag{1.2}$$

Crucially, the ith spin $\mathbf{s}_i$ in this expression has to be replaced with the masking token $\mathbf{t}$, since it is the masked input. The matrices $Q, K \in \mathbb{R}^{d\times d}$ are also trainable parameters of the model. In the following we take the embedding dimension equal to the number of colors, $d = C$, in order to be able to map the output vector $\mathbf{h}_i$ into a probability distribution $\tilde{\mathbf{p}}_i$ over the colors through the *softmax* non-linearity [1].

1.2.3 Training a transformer on the generalized Potts model

For our first experiment, we emulate the setting of protein structure prediction, so we choose a vocabulary of size $C = 20$ and sample a symmetric interaction matrix $J_{ij} = \{0, 1\}$ which we show in Fig. 1.2b). We draw the entries of the

[1] For a vector $\mathbf{x} = (x_i)$, the softmax non-linearity yields $x_i' = \exp(x_i)/\sum_j \exp(x_j)$. The elements of $\mathbf{x}'$ can be interpreted as a probability distribution, since they are positive and sum to one.

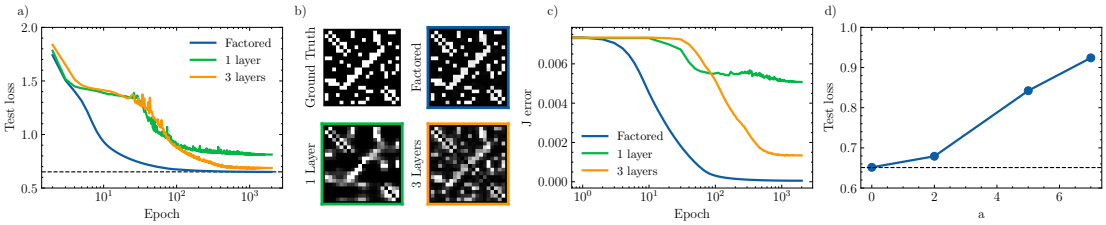


Figure 1.2: **A single layer of factored self-attention learns the generalized Potts model efficiently. (a)** Test loss [Eq. (1.3)] for factored self-attention and for vanilla transformers with one and three layers during training with stochastic gradient descent. The optimal generalization loss is shown as a black dashed line. **(b)** Interaction matrix J of the generative Potts Model [Eq. (1.1)] compared to the attention maps learned by transformers with vanilla and factored self-attention. For the three-layer transformer, the attention map was obtained by averaging the maps of the last two layers. **(c)** Reconstruction error of the interaction $(J - A)^2$ as a function of the number of epochs for all considered architectures. **(d)** Test loss as a function of perturbation level a. Decoupling the treatment between positions and colors by decreasing a decreases the test loss. *Parameters:* sequence length $L = 20$, vocabulary size $C = 20$, embedding dimension $d = 20$, $M = 3000$ data points.

symmetric interaction matrix U i.i.d. from the standard Gaussian distribution. Given these parameters, we use Gibbs sampling to generate a data set with $M = 3000$ sequences of length $L = 20$. We tune the inverse temperature β to ensure an average Hamming distance of 0.3 between sampled sequences, typical for protein families [69].

We then train off-the-shelf transformers consisting of one and three layers on this data set by minimising the cross-entropy loss between the output distribution and the missing spin using stochastic gradient descent on the loss $\mathcal{L}(\mathbf{s}) = -L^{-1} \sum_{i=1}^{L} \sum_{\alpha=1}^{C} s_{i\alpha} \log \tilde{p}_{i\alpha}(\mathbf{s})$, for a sequence $\mathbf{s}$. In Fig. 1.2, we show the test loss

$$\epsilon_g = \mathbb{E}_{\mathbf{s}\sim P} \left[\mathcal{L}(\mathbf{s}) \right] \tag{1.3}$$

during training, where $\mathbb{E}_{\mathbf{s}\sim P}$ denotes an average over the generative model (1.1). A transformer with a single layer of self-attention does not attain the optimal generalization error (black dashed line). By plotting the attention matrix of the single layer, we see that the transformer recovers the original interaction matrix to some degree, albeit not perfectly. Training transformers with three layers on the same data set improves the accuracy at the cost of loosing interpretability: there is no straightforward way to collapse several layers of non-linear transformations of the input sequence into a single attention map; we show the average of the final two attention layers in Fig. 1.2b.

1.2.4 Factored self-attention learns the generalized Potts model

We now consider a variant of self-attention in which the treatment of positions and values is decoupled. We set $a = 0$ in Eq. (1.2) , set the masking token $\mathbf{t} = 0$, choose one-hot encodings for the positions, and fix the embedding matrix at $E = I_C$, so that:

$$h_{i\alpha}\left(\mathbf{s}_{\setminus i}\right) = \sum_{j=1}^{L} A_{ij}(V\mathbf{s}_j)_\alpha \ , \tag{1.4}$$

where $A_{ij} \equiv e^{(Q^\top K)_{ij}} / \sum_{k=1}^{L} e^{(Q^\top K)_{ik}}$. This modified self-attention has exactly the same form as the conditional distribution of the generalized Potts model if one sets $U = V$ and $\beta J = A$, which is

$$p(s_{i\alpha} = 1|\mathbf{s}_{\setminus i}) = \frac{\exp\left(\beta \sum_{j=1}^{L} J_{ij}(U\mathbf{s}_j)_\alpha\right)}{\sum_{\gamma=1}^{C} \exp\left(\beta \sum_{j=1}^{L} J_{ij}(U\mathbf{s}_j)_\gamma\right)}. \tag{1.5}$$

This equivalence between factored self-attention and the Potts model is our first main result; we now discuss its ramifications.

Decoupling positions and colors leads to a significant improvement in the performance of a single layer, which reaches the optimal generalization error and converges faster, see Fig. 1.2a. Factored self-attention recovers the interaction matrix J perfectly, see Fig. 1.2b for the attention map and Fig. 1.2c for the reconstruction error of the interaction matrix. In Fig. 1.2d, we show that decoupling the treatment of positions and colors completely performs better than any intermediate solution with $a > 0$.

Factored attention layers, and thus input-independent attention weights, have been already used as a building block for deep transformers, outperforming standard attention in different applications: Ref.[70] used it to analyze protein sequences and found that a single layer of factored self-attention performed as well as a deep transformer, and significantly better than a single layer of vanilla self-attention, without explicitly explaining this observation. Moreover, using factored attention is key to obtaining state-of-the-art results in approximating ground states of many-body quantum systems, as we will see in detail in Chapter 4 [21, 37, 71].

Intriguingly, the form of the loss for masked language modeling with factored self-attention as described above exactly matches the loss of the pseudo-likelihood method, which has been used for solving the inverse Ising problem [44–48]. The pseudo-likelihood method is statistically consistent [8–10], i.e. its parameter estimates converge to the true parameters as the number of samples goes to infinity. A direct consequence of the mapping in Eq. (1.4) is thus that MLM with factored self-attention enjoys the same asymptotic optimality.

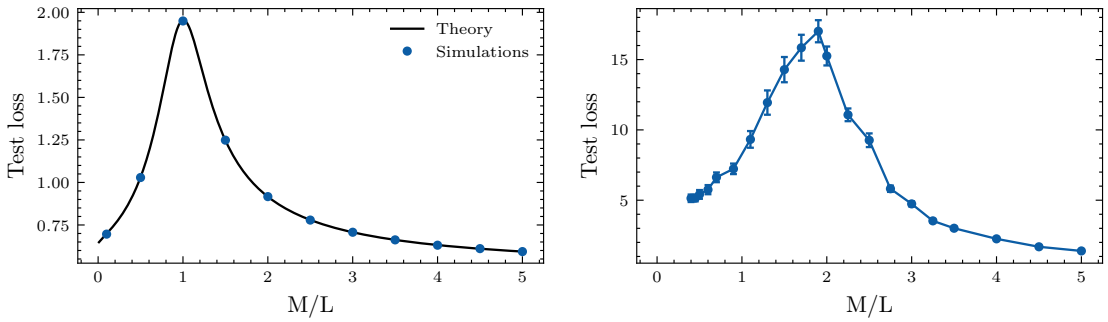


Figure 1.3: **The interpolation peak of factored attention in theory and practice. (Left)** A replica analysis predicts the test loss exactly. Test loss of a single layer of factored self-attention as a function of the number of samples per input dimension, as computed using replica theory (solid line). The blue points represent the outcome of numerical minimisation of the square loss [Eq. (1.6)], averaged over 30 realisations, and show perfect agreement with the theory. Error bars are smaller than point size. **(Right)** Same plot for a single layer of factored self-attention in the setting of Fig. 1.2 ($L = C = 20$), showing the same qualitative behaviour. The simulations are averaged over $n = 30$ different realisations.

1.3 The sample complexity of self-attention

A key quantity in machine learning problems is the sample complexity, namely how many samples are required to achieve a small generalization loss ϵ_g with a given model. The mapping introduced in this chapter allows us to address this question precisely for a single layer of self-attention by means of the replica method from statistical physics. The main difficulty in the calculation lies in handling the non-trivial data distribution [Eq. (1.1)]. This difficulty can be mitigated thanks to recent advances in statistical physics, which allow us to extend the replica method of disordered systems to structured data [61, 63, 72]. To perform the replica calculation, we first relax the discrete nature of Potts spins by re-writing the generalized Potts Hamiltonian [Eq. (1.1)] in terms of spin magnetization $\mathbf{m} = \langle \mathbf{s} \rangle_{P(\mathbf{s})}$ following mean-field theory. The associated Boltzmann measure then turns into a multivariate Gaussian distribution, whose covariance matrix is the negative inverse of the interaction matrix, i.e. $\Sigma = -J^{-1}$ [73, 74]. We then draw sequences $\{\mathbf{m}^\mu\}_{\mu=1}^M$ of length L from the multivariate Gaussian with zero mean and covariance matrix $\Sigma = \left(\Omega L^{-1/2} + \nu \mathbb{I}\right)^{-1}$, where Ω is a symmetric full-rank random matrix sampled from the Gaussian Orthogonal ensemble while $\nu\mathbb{I}$ is a diagonal matrix centering the spectrum of Ω in ν. To ensure Σ to be positive definite, we set $\nu > 2$ due to the semicircle law [75]. By fixing the location i of the masked spin across all input sequences, solving the MLM task is equivalent to inferring the ith row of

the interaction matrix $\mathbf{J}_i$.

To accomplish this task, we train a single layer of factored self-attention by empirical risk minimisation of a square loss with ℓ_2-regularisation:

$$\hat{\mathbf{A}}_i = \underset{\mathbf{A}_i}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{\mu=1}^{M} \left(m_i^\mu - \mathbf{A}_i \cdot \mathbf{m}_{\setminus i}^\mu \right)^2 + \frac{\lambda}{2} ||\mathbf{A}_i||_2^2 \right]. \tag{1.6}$$

Our goal is to characterise the generalization loss ϵ_g [Eq. (1.3)] of factored attention with the parameters obtained from minimising the loss in Eq. (1.6). In the high-dimensional limit, where the number of samples and the sequence length M, L tend to infinity while their ratio $\alpha \equiv M/L \sim O(1)$, we can express ϵ_g using the replica theory as a function of four scalar quantities:

$$\epsilon_g = \rho + q^\star - 2r^\star + 1/\nu. \tag{1.7}$$

Here, ν is the center of the spectrum of Ω, $\rho = \operatorname{tr} \Sigma_{\setminus i} / (\nu^2 L)$ is a function of the covariance matrix $\Sigma_{\setminus i}$ where we have removed the ith row and column, while $q^\star$ and $r^\star$ are the so-called overlap parameters. They correspond to practically measurable quantities over different realisation of the training set, involving the estimator of the ith row of the interaction matrix:

$$q^\star = \frac{\hat{\mathbf{A}}_i^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{L}, \qquad r^\star = -\frac{(\mathbf{A}_i^\star)^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{\nu L}. \tag{1.8}$$

The parameter r^* can be thus interpreted as an overlap between the estimated attention $\hat{\mathbf{A}}_i$ and the ground-truth value $\mathbf{A}_i$ while q^* is the overlap of the estimated attention, both mediated by the modified covariance matrix $\Sigma_{\setminus i}$.

As we show in Appendix 8.1.1, the values of these order parameters for a given training set of size α can be obtained by solving the optimisation problem

$$f_\beta = \underset{q,r,\delta q,\hat{q},\hat{r},\delta\hat{q}}{\operatorname{extr}} \left[-\frac{1}{2} (\hat{q}\delta q - q\delta\hat{q}) + \nu r\hat{r} + \lim_{L\to\infty} \frac{1}{L} \Psi_s + \alpha \Psi_e \right], \tag{1.9}$$

which yields the *typical* value (over the data) of the free energy density associated with a Gibbs measure at inverse temperature β whose Hamiltonian corresponds to the loss function in Eq. (1.6). Note that the optimisation only involves the scalar parameters q, r, δq and their conjugates $\hat{q}$, $\hat{r}$ and $\delta\hat{q}$, with $\delta q = q - q_s$, $\delta\hat{q} = \hat{q} + \hat{q}_s$ and q_s being the self-overlap among replicas. The so-called entropic potential Ψ_s is a function of the input covariance $\Sigma_{\setminus i}$:

$$\Psi_s = \lim_{L\to\infty} \frac{\hat{r}^2}{2L} \operatorname{tr} \left(\left(\Sigma_{\setminus i}\right)^t \Sigma_{\setminus i} \left(\nu\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i}\right)^{-1} \right) + \lim_{L\to\infty} \frac{\hat{q}}{2L} \operatorname{tr} \left(\Sigma_{\setminus i} \left(\nu\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i}\right)^{-1} \right), \tag{1.10}$$

while the energetic potential Ψ_e only depends on the specific choice of the loss function. As shown in Appendix 8.1.1, for the optimization problem in Eq. (1.6) Ψ_e is given by:

$$\Psi_e = -\frac{1}{2\,(1+\delta q)}\left(\frac{1}{\nu}+\rho+q-2r\right). \tag{1.11}$$

Using this approach, we estimated analytically the generalization loss in Eq. 1.7 as a function of the rescaled number of samples $\alpha = M/L$ (see Appendix 8.1.1 for details). The result is shown in the right panel of Fig. 1.3. As can be noticed, the test loss *increases* in the small data regime, before peaking at $\alpha = 1$. This value corresponds to the interpolation threshold, which is the largest number of samples that the neural network can perfectly fit, which in fact happens at $M = L$. Below this threshold, the model overfits to its training data; beyond this threshold, the generalization error decreases monotonically with the training set size; for large α, we found $\epsilon_g \sim \alpha^{-1/2}$. A similar peak in the generalization loss has been observed in supervised learning [76] and it is connected to the well-known "double descent" curve observed in deep neural networks in the presence of label noise [77]. There, the peak is a consequence of overfitting induced by the noise in the labels and it appears after an initial decay of the test loss at small α. In the self-supervised learning regime explored in this chapter, we find instead that the peak appears naturally as a consequence of the intrinsic stochasticity of the inputs. Indeed, in masked language modeling, the labels are a part of the input itself. The noise affecting the labels is thus highly correlated to that affecting the input. If the noise in the input is too high, the model starts immediately overfitting and the initial descent is not observed. The absence of the initial descent can be therefore ascribed to the high-level of the noise in the input.

We verify the predictions of the replica theory by plotting the generalization loss of a single layer of factored self-attention trained on the generalized Potts Model in the setting of Fig. 1.2 at small regularisation (right side of Fig. 1.3). We see the same qualitative behaviour as predicted by replica theory, even though in this case we did not apply any of the assumptions required for the replica analysis (the mean field limit and the usage of a full-rank J matrix and of a U matrix fixed to the identity). In particular, the test loss increases when adding more data for small α. The only difference between the plots is the location of the peak. For the square loss that we analyzed with replicas, as we already commented the peak is at the interpolation threshold $M = L$. For the simulations with logistic loss, the peak appears at the linearly separability threshold, which is the largest number of points a linear classifier can classify correctly, and which can be larger than one [54, 61].

1.4 Discussion

In this chapter, we have characterised the probability distributions that a single layer of self-attention can learn when trained on a masked language modeling task, considered as a simple prototype of self-supervised learning. In particular, we have shown analytically and numerically that with a single factored-attention layer, it is possible to exactly reconstruct the couplings of a generalized Potts model with two-body interactions between both sites and colors. More precisely, we showed that training factored self-attention on the MLM objective is equivalent to solving the inverse Potts problem using the pseudo-likelihood method [44–48], and therefore it yields consistent estimators of the parameters. These findings make factored attention a powerful, theoretically-driven building block for deep transformers. Our replica analysis of self-attention enabled us to compute the generalization loss of the model exactly and yielded a non-trivial generalization behaviour.

Chapter 2

A distributional simplicity bias in the learning dynamics of transformers

In this Chapter, we build on the results described in the previous Chapter on factored attention to investigate the presence of simplicity biases in transformers: transformer architectures trained on masked language modeling and next-token prediction tasks are able to learn the many-body interactions that govern natural text in a sequential way, going from lower to higher orders as training progresses. The main contributions of this chapter are the following:

- We introduce generative models based on transformers with factored attention, and show that the degree of interactions they capture is rigorously controlled by their depth (Section 2.2). The key element for generating the clones is the use of architectures combining factored attention layers developed in Chapter 1 with the square activation function. The depth of these architectures controls the degree of the many-body interactions captured;

- We study sequential learning of higher-order interactions both numerically and analytically in a controlled setting where we regulate the degree of interactions between tokens (Section 2.3);

- We train these generative models on benchmark natural language datasets, such as TinyStories [78], and sample them using Monte Carlo techniques to demonstrate the sequential learning of higher-order interactions by transformers in masked language modeling and next-token prediction (Section 2.4).

Our results show that the many-body interactions between tokens in natural text are key for learning, and we note that our approach naturally extends to other data modalities.

2.1 Simplicity biases in Transformers

The success of neural networks has been explained using various "simplicity biases", which postulate that neural networks trained using stochastic gradient descent (SGD) learn increasingly complex functions of their data, going from generalised linear models to more complex, non-linear functions [11–15]. These simplicity biases have been analysed theoretically in linear networks [79–81], in non-linear shallow neural networks [11, 82] and in the NTK regime [83–85]. The viability of such a scenario has also been demonstrated experimentally, for example in image classification with convolutional neural networks [14]. More recently, a complementary *distributional simplicity bias* was demonstrated theoretically and experimentally [86, 87]: deep neural networks learn increasingly complex approximations of the distribution of their training data. A natural question to ask is then: *Does this distributional simplicity bias extend to the realm of natural language processing (NLP), where transformers are typically pre-trained via Masked Language Modeling (MLM) or Next-Token Prediction?* As a first step in this direction, Belrose *et al.* [88] recently showed that deep transformers of the Pythia suite [89] first learn the token unigram and then the bigram distribution. However, estimating the higher-order interactions is prohibitively expensive on large corpora with many tokens due to the curse of dimensionality.

In this chapter, we design a novel framework based on MLM to build principled approximations of a NLP data set that captures increasingly higher-order interactions between tokens. Using these approximations or "clones", we show that deep transformers trained on natural language also exhibit a simplicity bias: they sequentially learn increasingly higher-order interactions among tokens [90].

The key idea of our framework is shown on the left of Fig. 2.1. Given a data set containing natural language (we will use WikiText-103 [94] as a running example), we first create a number of clones. Each clone approximates the underlying data distribution of WikiText-103, but only includes interactions between tokens up to a specified order (in the sketch, up to order three). We build these clones by using MLM to train a new type of transformer that consists of multiple layers of factored attention [91] with quadratic activation functions. In Section 2.2, we show analytically that the depth of this architecture precisely controls the degree of interactions among tokens in each clone. We then sample the trained models using state-of-the-art Monte Carlo techniques [95] to obtain the cloned data sets. The clones thus form a hierarchical family of data sets with increasing order of interactions among tokens that approximate the true distribution of WikiText-103 with increasing fidelity.

Using this set of clones, we investigate the sequential learning of higher-order interactions between tokens in a standard BERT-like transformer encoder [92, 93] with four attention blocks. On the right of Fig. 2.1, we show the test loss of this

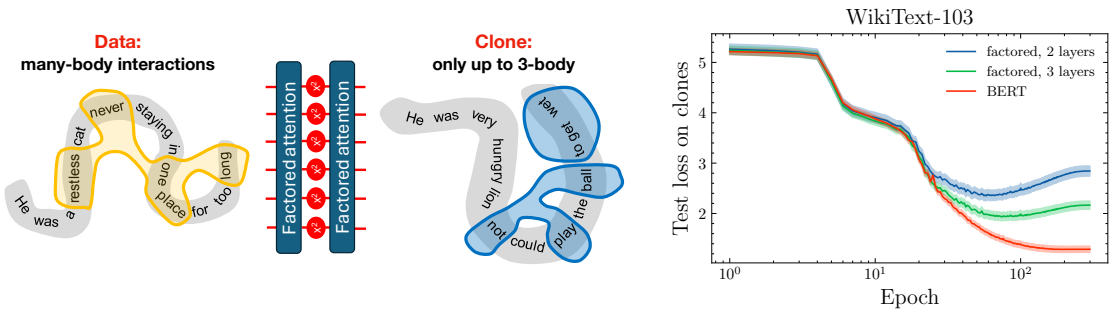


Figure 2.1: **Transformers learn increasingly higher-order interactions from their data. Left:** We illustrate the idea of a statistical "clone" of a data set, which approximates the underlying data distribution by keeping only interactions between tokens up to a fixed degree (in this case, three-body interactions). We introduce a principled approach to create clones by training a transformer with multiple layers of factored self-attention [91] with x^2 activation function between layers. The depth of the architecture controls the degree of the approximation. Clones can then be sampled from these models. **Right:** Test loss of a standard BERT-like transformer encoder [92, 93] with four attention blocks trained on the WikiText-103 [94] data set and tested on *clones* of this data set with a truncated maximum degree of many-body interactions between tokens. We show the average over five training runs starting from the same initial condition. The shaded area indicates one standard deviation.

architecture during MLM *training* on WikiText-103, but *testing* on various clones of WikiText-103 with a truncated maximum degree of many-body interactions between tokens, as well as on a clone sampled directly from the trained BERT model. The test losses of the transformer on the clones are close in the beginning of training up to ≈ 20 epochs. Beyond this threshold, BERT exhibits a different generalisation behaviour on clones that capture different orders of interaction: while the test loss on the clones containing only low-degree interactions among input tokens (blue and green curves) saturates after about 60 SGD epochs, the test loss evaluated on clones having high-order interactions (red curve) keeps improving during training. In other words, the accuracy of the transformer in the beginning of training derives from having learnt lower-order interactions, and only later during training the transformer also exploits higher-order interactions between tokens. In Section 2.4, we extend our experimental analysis to an alternative dataset, TinyStories [78], and further include experiments involving GPT models trained on next-token prediction tasks.

Further related work on simplicity biases in transformers Bhattamishra *et al.* [96] found a simplicity bias in the ability of transformers to learn sparse boolean functions; here, we focus on the MLM task on natural language. Abbe *et al.* [97] showed that the difference between trained and initial weights progressively increases in rank throughout training, hence increasing in complexity as the task requires. Vasudeva *et al.* [98] investigated a simplicity bias in transformers in terms of their sensitivity to random changes in the input. Other recent experimental work showed that RoBERTa achieves high performance on most linguistic benchmarks early in pre-training, while more complex tasks require longer pre-training time [99], and that transformers learn different linguistic phenomena in a prototypical way [100]. To the best of our knowledge, no prior research has investigated the progressive increase in complexity of the distribution learned by a transformer from its data over time. More recently, Cagnetta *et al.* [101, 102] studied probabilistic context-free grammars and showed that increasing the training set size allows transformers to learn correlations at increasingly large distances sequentially, corresponding to their resolving increasingly higher layers of the hierarchical structure of the tree. While we were preparing the camera-ready version of this paper, Garnier-Brun *et al.* [103] demonstrated a similar phenomenon in the training dynamics of transformers.

2.2 Expressing many-body distributions with factored attention

In this section, we present generative models utilising deep networks composed of multiple attention layers. We demonstrate analytically that these models, when trained with MLM, are capable of learning increasingly higher-order interactions among tokens. In the MLM task, transformers are trained to predict missing tokens in sentences of tokenized text. Following the same methodology introduced in Chapter 1, which we briefly summarize below, we model sentences as sequences of random variables $(s_1, \dots, s_L)$, where each token s_i is an element taking values in a discrete vocabulary: $s_i \in \{1, \dots, |\mathbb{V}|\}$. When masking the token in position i, the sequence $(s_1, \dots, s_i = 0, \dots, s_L)$ is given as input to the transformer, which embeds the input sequence into a collection of vectors $(x_1, \dots, x_L)$, each of them living in a d-dimensional space. Then, this sequence of vectors is processed by the self-attention layers and transformed into a sequence of context-dependent vectors $(y_1, \dots, y_L)$, where $y_j \in \mathbb{R}^d$ for all j. At the end, the output y_i, corresponding to the masked input token, is mapped into a vector of the same size of the vocabulary: $h_i \in \mathbb{R}^{|\mathbb{V}|}$. The prediction $p_{\text{mlm}}(s_i|s_{\neq i}) = \text{softmax}(h_i)$ can be finally interpreted as the probability distribution over all possible values the masked word can assume

within the vocabulary. The transformer is trained to minimise the MLM cross-entropy loss, $\mathcal{L} = -\sum_{m=1}^{M}\sum_{i\alpha} s_{i\alpha}^{(m)} \log\left(p_{\text{mlm}}(s_{i\alpha}^{(m)}|s_{j\neq i}^{(m)})\right)$, where M denotes the total number of examples in the dataset, and m indexes individual samples. The index i refers to the masked token position, while α represents the feature index. During training, the spins are cyclically masked one by one in each sequence.

The key ingredient to the transformer architecture [49] is the self-attention mechanism, which recombines input tokens into output tokens by weighing them with an input-dependent attention matrix. In Chapter 1, we considered the simplified attention mechanism called factored attention, where the attention weights are input-independent and the input sequence is processed as: $y_{i\alpha} = \sum_{j=1}^{L} A_{ij} \sum_{\beta=1}^{|\mathbb{V}|} V_{\alpha\beta} x_{j\beta}$, where $A \in \mathbb{R}^{L\times L}$ and $V \in \mathbb{R}^{d\times d}$ are matrices of trainable parameters. In this way, the prediction of a single factored attention layer for the masked token is expressed in terms of the following conditional distribution (assuming the embedding matrix to be the identity):

$$p_{\text{mlm}}(s_{i\alpha} = 1|s_{\neq i}) = \text{softmax}\left(\sum_{j\neq i}^{L} A_{ij} \sum_{\beta=1}^{|\mathbb{V}|} V_{\alpha\beta} s_{j\beta}\right) . \tag{2.1}$$

We have shown analytically that, when training a single layer of factored attention via MLM, the model is capable of exactly inferring the two-body interactions among input tokens, when sampled according to the joint distribution:

$$p(s) = \frac{1}{Z} \exp\left(-\sum_{i,j}\sum_{\alpha\beta} J_{ij} U_{\alpha\beta} s_{i\alpha} s_{j\beta}\right), \tag{2.2}$$

where J_{ij} describes the interactions among positions along the sequence, $U_{\alpha\beta}$ describes similarities between tokens and Z normalises the distribution. In the following, we introduce the convention of using Latin indices to run over the length of the sequences L and the Greek indices to run over the size of the vocabulary $|\mathbb{V}|$. As can be easily checked, the expression in Eq. (2.1) exactly matches the conditionals associated with the two-body joint distribution in Eq. (2.2).

We now demonstrate how we can easily express distributions with many-body interactions by stacking layers of factored self-attention. In particular, we introduce an architecture that can learn interactions of a specified order by tuning the number of layers. To this end, we start considering the simplest case where we stack just two layers of factored attention, with $f(x) = x^2$ as non-linearity between them and skip connections, see Fig. 2.2a) for a sketch of this architecture. By writing down the conditional distribution corresponding to this two-layer

model:

$$p(s_{i\alpha}=1|s_{j\neq i}) = \text{softmax}\left(\sum_{j\beta} A^{(2)}_{ij}V^{(2)}_{\alpha\beta}s_{j\beta} + \sum_{j\beta} A^{(2)}_{ij}V^{(2)}_{\alpha\beta}\sum_{k\gamma} A^{(1)}_{jk}V^{(1)}_{\beta\gamma}s_{k\gamma}\sum_{l\mu} A^{(1)}_{jl}V^{(1)}_{\beta\mu}s_{l\mu}\right), \tag{2.3}$$

we notice that the second layer expresses a three-body interaction, parameterised by a specific combination of the attention weights of the first and second layer through the tensor $\tilde{J}_{ikl} = \sum_j A^{(2)}_{ij}A^{(1)}_{jk}A^{(1)}_{jl}$ ($\tilde{U}$ is similarly expressed in terms of V). In particular, the first element of the sum in Eq. (2.3) accounts for the contribution of the skip connections, while the second one matches the functional form of the exact conditionals of a three-body joint distribution:

$$p(s_{i\alpha}=1|s_{j\neq i}) = \text{softmax}\left(3\sum_{jk}\sum_{\beta\gamma} J_{ijk}U_{\alpha\beta\gamma}s_{j\beta}s_{k\gamma}\right) . \tag{2.4}$$

Generalising these observations to a generic number of factored attention layers, we conclude that by stacking n layers of factored attention with x^2 activation function we can infer interactions up to order $2^{n-1}+1$. In this perspective, if the activation function is quadratic, the number of layers precisely controls the degree of interactions which can be inferred by a deep factored transformer.

2.3 Learning many-body interactions with factored attention

In this section we show how, in a controlled setup, the multi-layer factored attention model introduced in Section 2.2 sequentially learns increasingly higher-order interactions among tokens during training. To this end, we perform a numerical experiment where we sample the input sequences from a joint distribution with tokens interacting only up to the 4th-order, that is:

$$p\left(s\right) = \frac{1}{Z}\exp\left(-\sum_{ijkl}\sum_{\alpha\beta\gamma\mu} J_{ijkl}U_{\alpha\beta\gamma\mu}s_{i\alpha}s_{j\beta}s_{k\gamma}s_{l\mu}\right) . \tag{2.5}$$

Here, we ensure the couplings to be symmetric tensors by considering a set of vectors $\{j^p\}_{p=1}^P$ and setting $J_{ijkl} = \sum_{p=1}^P j^p_i j^p_j j^p_k j^p_l$. We choose $P = 200$ and we sample $j^p_i \overset{\text{iid}}{\sim} \text{Bernoulli}(0.8)$. The same holds for U. The exact conditionals corresponding to this joint distribution can be computed analytically, precisely as in the three-body case exemplified in Eq. (2.4). From the practical point of view, we use these conditionals to sample a set of M sequences $\{s^\mu\}_{\mu=1}^M$ via Gibbs sampling, which

will constitute the training set. We then train via MLM an increasing number of factored self-attention layers (from one to three) on these data, monitoring the test loss as a function of the number of epochs. The outcome of this experiment is shown in Fig. 2.2b). The dashed orange line corresponds to the lowest possible value of the test loss that a learning model would achieve in case it would be able to perfectly reconstruct the four-body tensors J and U of the generative model. The dashed blue and green lines correspond instead to the lowest test loss achievable by the best two and three body approximations of the input distribution, respectively. These last two bounds can be obtained by training via MLM a model whose predictions correspond to the exact conditional of a two and three-body distribution, where the learnable parameters are the tensors in Eq. (2.1) and Eq. (2.4). As shown by the red curve in Fig. 2.2b), if we train a single layer of factored attention on the four-body interacting sequences, the best we can do is to exactly recover the best two-body approximation [91]. By adding one extra layer of factored attention with x^2 non-linearity, we can reconstruct the best three-body approximation (violet curve). Then, the reconstruction of the full four-body interactions can be achieved by adding one further layer of factored attention (grey curve). This is consistent with the observation that by stacking n layers of factored attention with x^2 activation function we can infer interactions up to order $2^{n-1}+1$.

Figure 2.2b) also reveals that, in this controlled setup, the learning of higher-order interactions is preceded by distinct plateaus in the test loss. For example, in the three layers case, an initial plateau aligns with the best two-body approximation, followed by a subsequent plateau corresponding to the best three-body approximation. Interestingly, this sequential learning behaviour further implies that the layers are activated sequentially in time. This is illustrated in Figure 2.2c), where we display the Mean Square Displacement (MSD) of the weights for three layers of factored attention across training epochs, that is:

$$\mathrm{MSD}_l(t) = \frac{1}{L^2}||A^{(l)}(t) - A^{(l)}(0)||^2 + \frac{1}{|\mathbb{V}|^2}||V^{(l)}(t) - V^{(l)}(0)||^2 \tag{2.6}$$

with $A^{(l)}(t)$ and $V^{(l)}(t)$ being the attention and the value matrix of the l-th layer at time t. As we can see, despite the simultaneous update of all network parameters during training, a distinctive sequential learning pattern emerges. In the early epochs, only the last layer (orange curve) activates concurrently to the appearance of the first plateau, thus suggesting that the last layer is the one responsible for learning the best two-body fit of the training data. Subsequently, the weights of the second layer come into play (green curve), and combined with the weights of the last layer can reconstruct the best three-body approximation of the underlying four-body data distribution. Close to the end of the training, also the weights of the first layer start changing and, combined with the other layers, effectively infer the ground truth four-body interaction among tokens in the training data. A hint

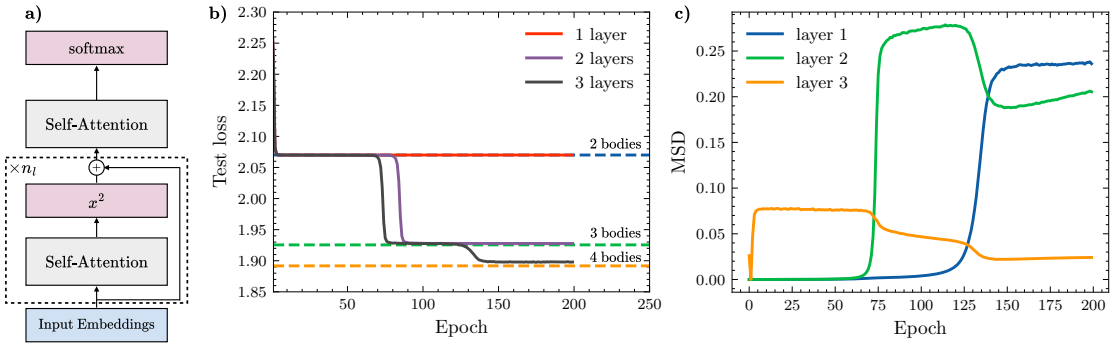


Figure 2.2: **a)** Multi-layer factored self-attention architecture with x^2 activation function. **b)** Test loss learning curves of one, two and three factored self-attention layers with x^2 activation function. The models were trained on a synthetic data set generated from a four-body Hamiltonian. The dashed horizontal lines correspond to the convergence value of the loss for two, three and four bodies energy based models trained on the same data set. **c)** Mean Square Displacement of the weights across different layers in a three-layers factored attention architecture. In these experiments, the size of the vocabulary was set to $|\mathbb{V}| = 10$ and the sequence length to $L = 20$. We used a training set of $M = 25600$ samples, training the models with SGD, choosing a mini-batch size of 256. The initial learning rate is chosen to be 0.1.

of the reason why the weights of the last layer find the best two-body interaction is that, due to the presence of skip connections, they appear linearly combined with the input tokens in the output prediction. In the next section, we provide a more formal analysis of this phenomenon.

2.3.1 Analysing the gradient flow for two-layer factored attention

To gain theoretical insights on the sequential learning behaviour observed in the previous numerical experiments, we analyse the gradient flow of two-layers of factored attention with x^2 activation function when trained via MLM. For the sake of simplicity, we restrict to the case $d = 1$, that is when the tokens in the sequence $(x_0, x_1, \dots, x_L)$ are real numbers rather than vectors. We consider the learning framework where the model is trained to always predict the first element of each sequence, using the square loss $\mathcal{L} = \frac{1}{2}(x_0 - y_0)^2$. Here, y_0 represents the transformer prediction about the masked token which, in this setting, acquires the following

functional form:

$$y_0 = \sum_j A^{(2)}_{0j} x_j + \sum_j \sum_k \sum_l A^{(2)}_{0j} A^{(1)}_{jk} A^{(1)}_{jl} x_k x_l \ . \tag{2.7}$$

The gradient flow equations can then be written explicitly for the attention weights of both factored attention layers as:

$$\dot{A}^{(2)}_{0j} = \left(x_0 - \sum_j A^{(2)}_{0j} x_j - \sum_j A^{(2)}_{0j} \lambda_j^2 \right) \left(x_j + \lambda_j^2 \right) \tag{2.8}$$

$$\dot{A}^{(1)}_{kl} = 2 \left(x_0 - \sum_j A^{(2)}_{0j} x_j - \sum_j A^{(2)}_{0j} \lambda_j^2 \right) \left(A^{(2)}_{0k} x_l \lambda_k \right) , \tag{2.9}$$

where $\lambda_k = \sum_l A^{(1)}_{kl} x_l$. From these equations, we can immediately see that, if we initialise the network with zero weights $A^{(1)} = A^{(2)} = 0$, the first layer attention weights do not change in time, $\dot{A}^{(1)}_{kl} = 0$, while the second-layer weights do, $\dot{A}^{(2)}_{0j} \neq 0$, consistently with what we observe in Fig. 2.2c).

We can gain some further intuition into which degree of interactions each layer learns by simplifying the data model further. In particular, we can restrict to a Gaussian design where tokens are now drawn from a zero-centered Gaussian distribution. By taking the expectation over the data, the correlations of an odd number of random variables vanish and the gradient flow equations can be written as:

$$\dot{A}^{(2)}_{0j} = h_j - \left(C A^{(2)}_0 \right)_j - \sum_k A^{(2)}_{0k} \mathbb{E} \left[\lambda_j^2 \lambda_k^2 \right] \tag{2.10}$$

$$\dot{A}^{(1)}_{kl} = -2 A^{(2)}_{0k} \sum_j A^{(2)}_{0j} \mathbb{E} \left[\lambda_j^2 x_l \lambda_k \right] \ , \tag{2.11}$$

where we have defined $h_j = \mathbb{E}[x_0 x_j]$ for $j > 0$ and $C_{ij} = \mathbb{E}[x_i x_j]$ for $i, j > 0$. In Appendix 8.2.1 we further show that, if data are Gaussian distributed, the expectation value of the loss is bounded from below by the value it takes for $A^{(1)} = 0$. Therefore, if one adds an ℓ_2 regularisation of arbitrarily small strength, $A^{(1)} = 0$ is the only stable fixed point of the gradient flow equations. This is consistent with the numerical experiments, which show that for two-body distributed data the weights of the first layer vanish during the optimisation. In this case, only the second layer weights contribute to the learning dynamics:

$$\dot{A}^{(2)}_{0j} = h_j - \left(C A^{(2)}_0 \right)_j \ , \tag{2.12}$$

whose solution describes an exponential convergence in time to $A^{(2)}_{0j}(t \to \infty) = (C^{-1} h)_j$. This stationary point reproduces the exact conditional distribution of a

Gaussian entry given all the others. Indeed, considering a multivariate Gaussian distribution with zero mean, indicating the first row of the covariance matrix (except the $0,0$ element) by h and the block obtained with $i,j>0$ by C we have: $p(x_0|x_{>0}) = \mathcal{N}(x_{>0}^T C^{-1} h, \Sigma)$. This means that the second layer is learning the covariance matrix, and thus the interactions up to the second order.

If data contains higher-order interactions among tokens, we need also the first layer to unveil the underlying data distributions. Once again, we can see this from the gradient flow equations. In the general case of higher-order interactions (thus for non-Gaussian data distributions), the stationary point $A(t)=0$ is generally unstable. For example, considering the case of a diagonal perturbation $A^{(1)}_{kh} = \epsilon \delta_{kh}$ around $A^{(1)}=0$, Eq. (2.9) becomes (neglecting terms of order ϵ^2):

$$\dot{\epsilon}\delta_{kl} = \epsilon A^{(2)}_{0k} \left(\Gamma_{0lk} - \sum_j A^{(2)}_{0j} \Gamma_{jlk} \right) , \tag{2.13}$$

where we have taken the expectation with respect to a generic data distribution and defined $\Gamma_{ijk} = \mathbb{E}\left[x_i x_j x_k\right]$. In this case $A^{(1)}(t)=0$ is a stable solution with respect to a diagonal perturbation only if the matrix $G_{lk} = A^{(2)}_{0k}\left(\Gamma_{0lk} - \sum_j A^{(2)}_{0j}\Gamma_{jlk}\right)$ is negative definite (implying that the perturbation decays to zero as t goes to infinity). Therefore, the structure of the three-body correlations Γ_{ijk} determines the stability of the stationary point $A^{(1)}=0$. Indeed, if, for example, one takes $\Gamma_{jlk} = \gamma \delta_{ij}\delta_{jk}$, we have $G_{lk} = \gamma \delta_{lk} A^{(2)}_{0k}\left(\delta_{0k} - A^{(2)}_{0k}\right)$, which is not negative definite for a generic choice of $A^{(2)}_0$. This analysis allows concluding that there are conditions for which the solution $A^{(1)}=0$ is unstable when data contains interactions of order higher than second, which is precisely in line with what we observe in Fig. 2.2b).

2.4 Sequential learning in NLP

In this section, we use the tools introduced in the previous sections to investigate which features of the input data influence the training dynamics of transformers pretrained on real NLP datasets. Specifically, we analyze the TinyStories dataset [78], training a standard BERT encoder [92] using the masked language modeling objective, as well as a standard GPT model for the next-token prediction task. Our focus lies on understanding whether a sequential learning of many-body interactions among input tokens, akin to the observed phenomenon in the synthetic scenario, is also prevalent in a real-world context. Since we do not know the underlying joint distribution over words, our approach will be to first train the attention-based models of Section 2.2 with varying numbers of layers on

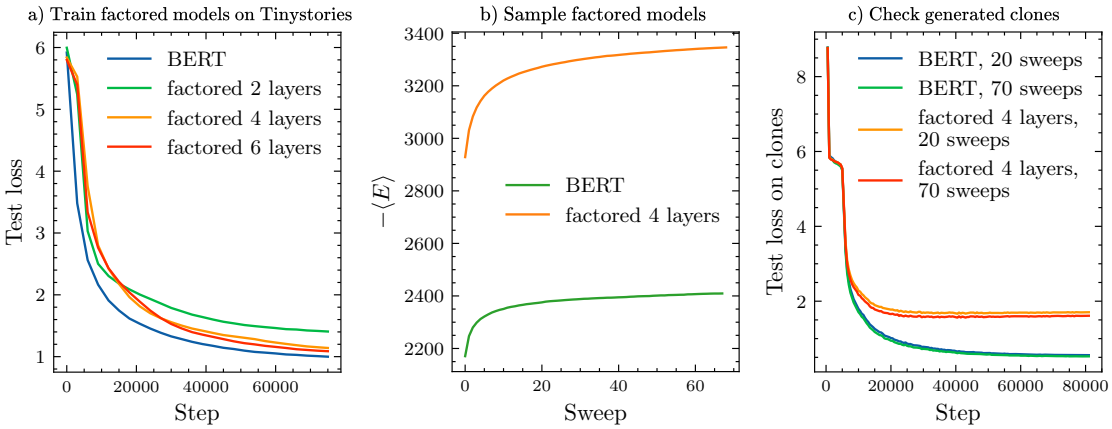


Figure 2.3: **Three steps for cloning a data set using factored-attention based generative models.**
a) Train factored-attention models on TinyStories. Test loss curves of different factored-attention based architectures trained on TinyStories and tested on TinyStories. Specifically, we consider architectures with two, four and six factored self-attention layers with x^2 activation function. For comparison, also the test loss of a four-layers BERT is shown. **b) Sample factored models.** Mean score of a batch of sentences taken from the test set of the TinyStories data set and evolved with the Metropolis-Hasting sampling scheme described in Section 8.2.2. **c) Check generated clones.** Test loss curves of a standard four layers transformer encoder, trained on TinyStories and tested on clones generated after 20 and 70 Metropolis-Hasting sweeps. The clones were generated from a four layers standard BERT and from an architecture with four layers of factored self-attention and x^2 activation function (associated with a nine bodies approximation of TinyStories).

TinyStories, that can be sampled to generate "clones" of the original data with a truncated maximum degree of interaction. We now discuss our experimental methods and analyze the results of our experiments; for full details on the methods, see Appendix 8.2.2.

The clones We exploit the generative models described in Section 2.2 to generate clones of the original TinyStories data set (see Appendix 8.2.2 for details on how we prepare the data set). The clones are meant to approximate the underlying TinyStories distribution with increasing fidelity. We use $n = 2, 4$ and 6 layers of factored attention to generate clones that include effective interactions up to

Original text	Once upon a time there lived a cat named Max. He was a restless cat, never staying in one place for too long. One day, Max demanded something funny. He wanted a friend.
Factored 2 layers	Once upon a time there was a boy named Timmy. He was very hungry lion the with not could play the ball to get wet. One day, Lily saw something else. He shook his head.
Factored 4 layers	Once upon a time there was a boy named Tom. He saw a small cat, could not play in the park with flowers pretty. One day, Lily smelled and nice. They wanted the ball.
BERT	Once upon a time there was a puppy named Max. Max was a small puppy who was buzzing in a basket for very long. One day, Max saw something amazing. He wanted a puppy.

Table 2.1: **Sampling the clones.** In the first row, we show part of a sentence taken from the test set of TinyStories. The second, third and fourth rows show how the original text is modified after 20 sweeps of Monte Carlo sampling associated to two and four layers factored architectures and BERT architectures, respectively.

degree 3, 9, and 33, respectively. Since we use a vocabulary of size $|\mathbb{V}| = 10000$, we perform a *finite*-rank approximation of the value matrix with rank 400 (see Appendix 8.2.2 for more details). In Fig. 2.3a), we show the test loss curves when training models with two, four, and six layers of factored attention featuring the x^2 activation function on TinyStories. Notably, the final value of the test loss decreases systematically with the order of the interaction modeled by the architecture; for comparison, the test loss curve associated with the training of a *standard* BERT architecture [92] is also shown.

Sampling procedure The bidirectional architectures we consider here are more expressive, yet also more challenging to sample than auto-regressive models such as GPT [104], as we discuss with an example in Appendix 8.2.2. Here, we adopt the sampling approach of Goyal *et al.* [95] which is based on the Metropolis-Hastings (MH) algorithm. We first select 620 examples of the test set at random. For each input sequence $(s_1, \ldots, s_L)$, we define its score in terms of the masked conditionals used for training the masked language models, $E(s_1, \ldots, s_L) = -\sum_{i=1}^{L} h_i \cdot s_i$, with h_i being the logits as defined in Section 2.2. Then, a MH sampler is run on the joint distribution defined as $p(s_1, \ldots, s_L) \propto \exp\left[-E(s_1, \ldots, s_L)\right]$. Starting from an initial sequence $(s_1^{t=0}, \ldots, s_L^{t=0})$, at each step a local move is proposed, in which the i-th token is masked and replaced with $s_i^{t+1} \sim p_{\text{mlm}}(s_i | s_{\neq i}^t)$. The move is then accepted according to the standard MH criterion. In this framework, the conditional distributions obtained after training the transformer with MLM

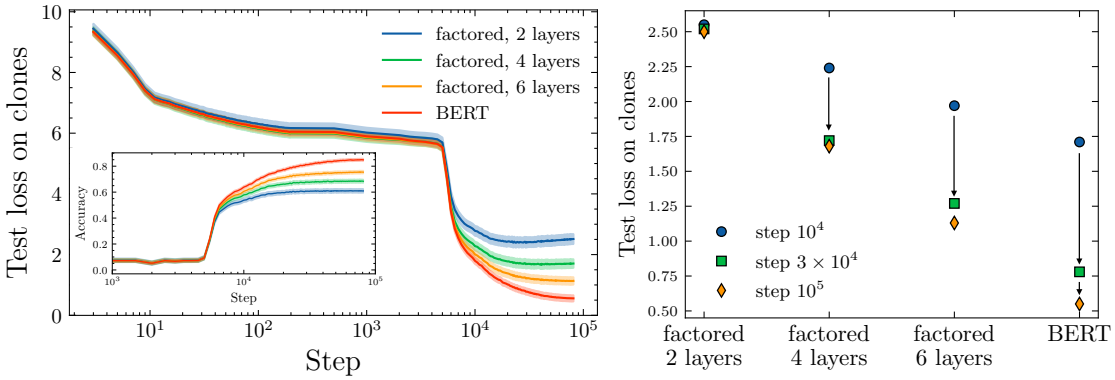


Figure 2.4: **BERT models trained on masked-language modelling learn increasingly higher-order interactions during training. Left panel:** In an experiment analogous to the one shown in Fig. 2.1, we show the test loss of a standard BERT-like transformer encoder trained trained on the TinyStories data set [78] and tested on *clones* of this data set with a truncated maximum degree of many-body interactions between tokens. The inset shows the corresponding test accuracy. We show the average over five different training runs, all starting from the same initial condition. The shaded area indicates one standard deviation. **Right panel:** An alternative way to visualise the data from the left panel is to plot the test loss at steps 10^4, 3×10^4, and 10^5 (blue, green and orange points respectively). This visualisation highlights the sequential learning of higher-order interactions, showing that for the clones derived from two- and four-layer factored architectures the loss saturates after 3×10^4 training steps, while on the clones derived from a six-layer architecture, as well as for the clone sampled from a BERT model, the test loss continues to decrease, as indicated by the black arrows.

are used as the transition kernel of the MH algorithm. In Fig. 2.3b), we show the evolution of the mean score after each sweep of the sampling procedure, in the case of the standard BERT transformer and the four layers factored architecture. In Table 2.1, we finally show how one randomly selected part of a story from the test set of TinyStories is modified after 20 sweeps of Monte Carlo sampling associated to two and four layers of factored attention, and four layers BERT architecture.

Validating the clones As a sanity check of the statistical consistency of the clones, we trained a standard four-layer transformer encoder on TinyStories, but tested it on the clones generated after 20 and 70 MH sweeps, see panel c) of Fig. 2.3. For both the clones generated from a four-layer factored architecture and from a standard four-layer transformer encoder, we can check that the test loss after 20 and 70 sweeps is really similar, thus we decided to consider the samples obtained

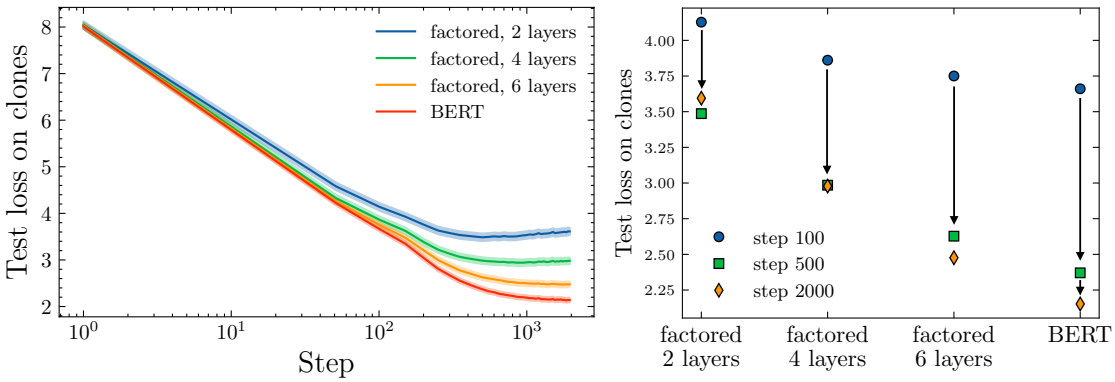


Figure 2.5: **GPT models trained on next token prediction tasks learn increasingly higher-order interactions during training. Left panel:** Test loss of a two-layer GPT-Neo model from [78] trained on TinyStories set and tested on clones of this data set with a truncated maximum degree of many-body interactions between tokens. **Right panel:** We repeat the analysis of Fig. 2.4 for the GPT-Neo model trained using next-token prediction, showing again clear signatures of sequential learning. The results are consistent with respect to different random seeds.

after 20 sweeps reliable clones of the original data set.

Sequential learning of many-body interactions by BERT As already anticipated, we use these clones to investigate how a standard transformer encoder trained on MLM learns the different degrees of interactions among input words. Our results in Fig. 2.1 already suggested a sequential learning of the various degrees of interaction among words over time for the WikiText-103 dataset. In Fig. 2.4, we show an additional experiment where we trained the same transformer (for more details about the architecture see Section 8.2.2) on a different data set, Tinystories [78], finding the same sequential learning dynamics. We show this effect in the left panel of Fig. 2.4, where we plot the test loss of the four-layer BERT model of Fig. 2.1 *trained* on TinyStories, while *tested* on clones with a fixed maximum degree of interactions that we generated. After the initial plateau, which finishes at $\sim$ 5000 training steps, the loss curves on the various clones continue to overlap up to $\sim$ 6000 steps, despite the quick improvement in test error. This behaviour suggests that in the early stages of training, the model primarily learns low-order interactions without yet capturing high-order ones. Furthermore, in the final phase of training, the test error on the three-body clone, derived from a two-layer factored architecture, shows little change after 3×10^4 steps (blue curve). A similar behaviour happens for the clone obtained from a four-layer architecture (green curve). Instead, on the clone obtained from a six-layer factored architec-

ture and from a BERT architecture, the loss continues to decrease (orange and red curves respectively) until the end of the training process, highlighting the sequential learning of the higher-order interactions by BERT. In the right panel of Fig. 2.4, we present an alternative visualization of the same test loss values on the clones at steps 10^4, 3×10^4, and 10^5 (blue, green and orange points respectively). This visualization highlights that, for the clone derived from a two-layer factored architecture, training beyond 10^4 steps does not change the test error significantly. In contrast, for the clone based on a four-layer factored architecture, the test loss shows significant improvement between 10^4 and 3×10^4 steps, but effectively saturates after 3×10^4 steps. Notably, for the clone derived from a six-layer architecture, as well as for the clone sampled from a BERT model, the test loss continues to decrease throughout the entire training process, as indicated by the black arrows.

Sequential learning of many-body interactions by GPT Finally, we extend the analysis to the learning dynamics of an autoregressive transformer model trained for the next-token prediction task. Specifically, we examine a two-layer GPT-Neo model, as also investigated in the TinyStories paper [78], training it on the original dataset and evaluating its performance on the different clones (see Appendix 8.2.2 for full experimental details). While the loss curve shows a single decay of the loss, the sequential learning is again clearly apparent, as shown in the left panel of Fig. 2.5. An alternative visualization of the same data presented in the left panel of the same figure reveals that the test loss on clones with low-order interactions saturates within the first 500 training steps, while the test loss on clones describing higher-order interactions continues to decrease throughout the entire training period. Crucially, these findings demonstrate the applicability of our framework also to autoregressive models, which are the foundation of state-of-the-art large language models.

2.5 Discussion

In this chapter, we have shown that transformer architectures trained on masked language modeling and next-token prediction tasks are able to learn the many-body interactions that govern natural text in a sequential way, going from lower to higher orders as training progresses. We showed this using a new approach to assess simplicity biases by first generating clones of a NLP data set, where the maximum level of interaction among tokens is limited to a specific order. Testing a transformer on these clones while training it on the original data set provides insights into the sequential learning patterns of the transformer. We conducted numerical simulations to show that, when training a transformer on the WikiText-103 and

TinyStories data sets, the test loss curves for clones with low-degree interactions (e.g., three-body) exhibit plateaus after a fraction of the total training steps. In contrast, higher-order interactions exhibit continuous test loss reduction throughout the entire training procedure. The key element for generating the clones is the use of architectures combining factored attention layers (see Chapter 1) with the square activation function; the depth of these architectures controls the degree of the many-body interactions captured.

Chapter 3

Vision Transformer Wave Function

In this Chapter, we build on the factored attention mechanism introduced in Chapter 1 to develop an adaptation of the Vision Transformer (ViT) architecture [105] capable of generating a variational ansatz of quantum states for many-body systems. We refer to this ansatz as the ViT wave function. Following the most used approach in the literature of Neural Quantum states [17, 19], we generalize the ViT architecture by introducing complex-valued parameters into the network. In this specific application, the factored attention mechanism is crucial for achieving accurate results, demonstrating superior performance compared to the standard attention mechanism used in natural language processing (a detailed comparison will be presented in Chapter 6). We use the developed architecture to study 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 shallow architecture, with a *single layer of factored self-attention*, thus largely simplifying the original Vision Transformer architecture [105]. 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. Nevertheless, the presence of complex parameters in the network complicates the optimization of deep architectures, thereby motivating the introduction of a dedicated training protocol to enhance optimization stability.

3.1 Introduction to Neural Quantum States

Variational approaches for studying quantum many-body systems have proved fundamental for understanding the ground state properties of extremely complicated physical systems, famous examples being the Bardeen-Cooper-Schrieffer state [106]

and Laughlin [107] wave functions to explain superconductivity and the 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 non-trivial task for strongly-interacting systems. An approximate technique to address this problem is the Variational Monte Carlo (VMC) method, which involves minimizing the variational energy $E_\theta = \langle\Psi_\theta|\hat{H}|\Psi_\theta\rangle \,/\, \langle\Psi_\theta|\Psi_\theta\rangle$, where $|\Psi_\theta\rangle$ is a variational state parametrized through a vector $\boldsymbol{\theta}$ of P parameters [16]. For a system of N 1/2-spins, $|\Psi_\theta\rangle$ can be expanded in the computational basis $|\Psi_\theta\rangle = \sum_{\{\sigma\}} \Psi_\theta(\sigma)\,|\sigma\rangle$, where $\Psi_\theta(\sigma) = \langle\sigma|\Psi_\theta\rangle$ is a map from spin configurations of the Hilbert space, $\{|\sigma\rangle = |\sigma_1^z, \sigma_2^z, \cdots, \sigma_N^z\rangle\,,\ \sigma_i^z = \pm 1\}$, to complex numbers. In a gradient-based optimization approach, the fundamental ingredient is the evaluation of the gradient of the loss, which in this case is the variational energy, with respect to the parameters θ_α for $\alpha = 1, \ldots, P$. This gradient can be expressed as a correlation function [16]

$$F_\alpha = -\frac{\partial E_\theta}{\partial \theta_\alpha} = -2\,\mathrm{Re}\left[\langle(\hat{H} - \langle\hat{H}\rangle)(\hat{O}_\alpha - \langle\hat{O}_\alpha\rangle)\rangle\right]\ , \tag{3.1}$$

which involves the diagonal operator $\hat{O}_\alpha$ defined as $O_\alpha(\sigma) = \partial \mathrm{Log}[\Psi_\theta(\sigma)]/\partial\theta_\alpha$. The expectation values $\langle \ldots \rangle$ are computed with respect to the variational state and stochastically estimated by running a Markov Chain on the distribution $P_\theta(\sigma) \propto |\Psi_\theta(\sigma)|^2$, see Appendix 8.3.1 for more details. The parameters are then updated according to the Stochastic Reconfiguration (SR) method [16, 108, 109], which takes into account the geometric structure of the landscape:

$$\boldsymbol{\delta\theta} = \tau\,(S + \lambda\mathbb{I}_P)^{-1}\,\boldsymbol{F}\ , \tag{3.2}$$

where τ is the learning rate and λ is a regularization parameter to ensure the invertibility of the S matrix. The matrix S has shape $P \times P$ and it is defined in terms of the $\hat{O}_\alpha$ operators [16]

$$S_{\alpha,\beta} = \mathrm{Re}\left[\langle(\hat{O}_\alpha - \langle\hat{O}_\alpha\rangle)^\dagger(\hat{O}_\beta - \langle\hat{O}_\beta\rangle)\rangle\right]\ . \tag{3.3}$$

The choice of the parametrization for the variational wave function $|\Psi_\theta\rangle$ is a central component of the VMC method. Recently, a class of wave functions, based on neural networks, has been introduced and developed [17, 18]. Starting from Restricted Boltzmann Machines (RBMs) [17], 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) [29–32], Recurrent Neural Networks (RNNs) [33, 34], Autoregressive Neural Networks [35, 36], but also combinations

of neural networks with standard variational wave functions (e.g., Gutzwiller-projected fermionic ones) [110, 111]. In the last few years, the Transformer architecture [1] 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 so-called Vision Transformer (ViT) [105], 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 [35], to perform quantum tomography in presence of noise [112], and for real- and imaginary-time evolutions of quantum systems [113].

In this Chapter, 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. Crucially, the factored version of the self-attention mechanism developed in Chapter 1 is essential to reach accurate results. We apply our *Ansatz* to the one-dimensional J_1-J_2 Heisenberg model, whose Hamiltonian is defined by

$$\hat{H} = J_1 \sum_R \hat{\boldsymbol{S}}_R \cdot \hat{\boldsymbol{S}}_{R+1} + J_2 \sum_R \hat{\boldsymbol{S}}_R \cdot \hat{\boldsymbol{S}}_{R+2} \tag{3.4}$$

where $\hat{\boldsymbol{S}}_R = (S^x_R, S^y_R, S^z_R)$ 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 [114]. 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 two-fold 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 $(J_2/J_1)_c = 0.241167 \pm 0.000005$ [115, 116]. 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) [117] or its modern variations based upon tensor networks *Ansätze* [118] 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. (3.4). In fact, the main limitation to the use of quantum Monte Carlo techniques [119] 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 [120]). The non-trivial sign structure represents also an obstacle to the definition of accurate variational wave functions. For example, Gutzwiller-projected fermionic states [121] have a limited power to reproduce the correct signs of the ground state for $J_2/J_1 > 0.5$ [122]. 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 [122]. 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). RNNs *Ansätze* have also been 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 [123].

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 to mix 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 Chapter, we show that the ViT wave function can reach very high accuracy results compared to DMRG calculations, even on large clusters, with less than one thousand parameters and few computational resources compared to other neural-network wave functions. Most importantly, the ViT accuracy can be systematically improved by changing the hyper-parameters of the architecture.

3.2 The ViT wave function

The fundamental ingredient of a Transformer is the *self-attention mechanism*. As we saw in Section 1.2, given a sequence of N input vectors $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$, for each of them three new vectors are computed, $\boldsymbol{q}_i = \mathrm{Q}\boldsymbol{x}_i$, $\boldsymbol{k}_i = \mathrm{K}\boldsymbol{x}_i$, and $\boldsymbol{v}_i = \mathrm{V}\boldsymbol{x}_i$, 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 j-th input vector should contribute to $\boldsymbol{A}_i$, which is the subsequent representation of the i-th input. The functional form of these weights can be chosen according to the task [2]. To improve the performance of the model, multi-head attention can be considered, where a set of matrices Q^μ, K^μ, and V^μ, with $\mu = 1, \dots, h$ (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

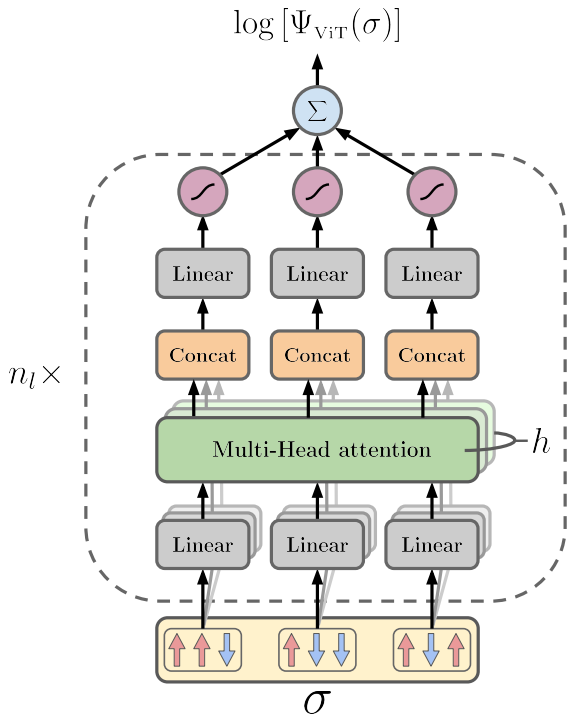


Figure 3.1: Scheme of the ViT wave function. The input spin configuration σ is split into patches of size b (which define a set of N vectors of dimension b). Each of them is linearly projected h times with different linear projections to produce N vectors of dimensions $r = d/h$. Then the attention function 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 $\text{Concat}(\boldsymbol{A}_i^1, \dots, \boldsymbol{A}_i^h)$ 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.

together, and linearly combined. Finally, each output vector of the multi-head attention is fed separately and identically to a non linearity. In general, this whole architecture is replicated n_l times.

Our goal is to use the Transformer to parameterize the many-body wave function, in order to map spin configurations of the Hilbert space $\sigma = (\sigma_1, \dots, \sigma_L)$, with $\sigma_R = 2S_R^z = \pm 1$, to complex numbers $\Psi(\sigma)$. We take inspiration from the ViT [105] 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 $\sigma = (\sigma_1, \dots, \sigma_L)$, we split it into N patches of b elements: $\boldsymbol{x}_i = (\sigma_{(i-1)b+1}, \dots, \sigma_{(i-1)b+b})$, for $i = 1, \dots, N$ (the total number of sites must be a multiple of b). The sequence of these patches is then used to

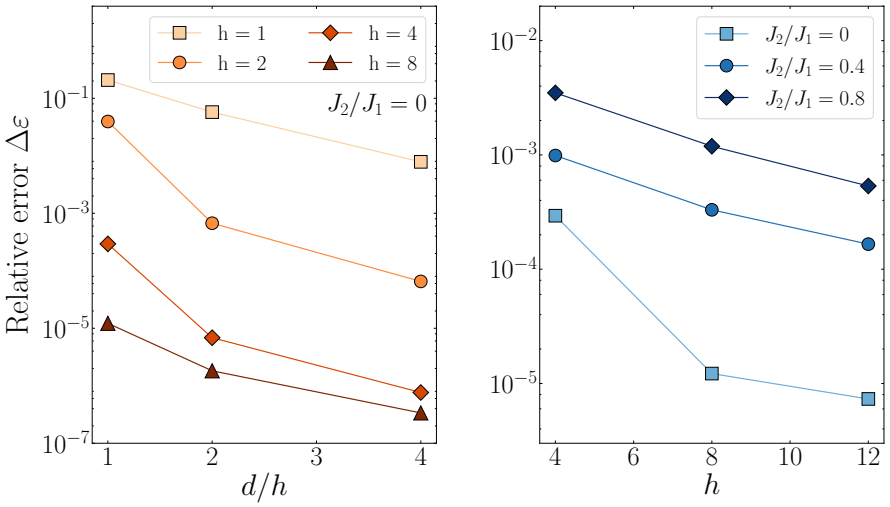


Figure 3.2: Relative error $\Delta\varepsilon = |(\mathrm{E_{ViT}} - \mathrm{E_{DMRG}})/\mathrm{E_{DMRG}}|$ of the ViT wave function by varying the hyper-parameters of the architecture for a cluster with $L = 100$ sites. Left panel: $\Delta\varepsilon$ as a function of d/h, with a fixed number of heads 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 [124] 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 $J_2/J_1 = 0.8$.

compute the attention vectors. Then, the factored attention version of the original self-attention mechanism developed in Chapter 1 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:

$$\boldsymbol{A}_i^\mu = \sum_{j=1}^{N} \alpha_{ij}^\mu \mathrm{V}^\mu \boldsymbol{x}_j, \tag{3.5}$$

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. In this specific application, 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 non linearity, 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. 3.1).

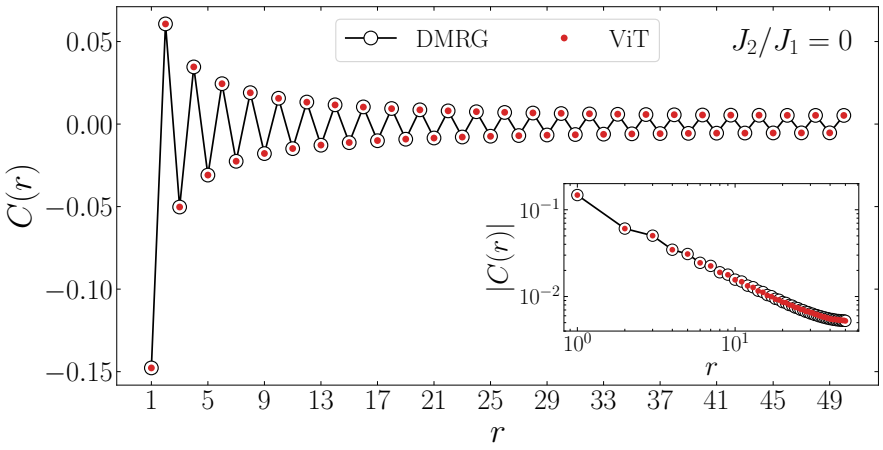


Figure 3.3: The isotropic spin-spin correlations in real space $C(r)$ as computed by the ViT wave function (full dots) for the unfrustrated Heisenberg model ($J_2/J_1 = 0$) on a cluster with $L = 100$ sites. The DMRG results are also shown for comparison (empty circles). Inset: Log-log plot of the same correlation function.

In order to study frustrated quantum spin models with non-positive ground states (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* [125] to periodic systems, taking $\alpha^{\mu}_{i,j} = \alpha^{\mu}_{i-j}$; as a result, the number of variational parameters for computing the attention vectors (3.5) is reduced from $O(L^2)$ to $O(L)$. This procedure induces translational invariance between patches. To include also the one within patches, we perform the linear combination:

$$\tilde{\Psi}_{\mathrm{ViT}}(\sigma) = \sum_{r=0}^{b-1} \Psi_{\mathrm{ViT}}(\mathrm{T}_r \sigma), \tag{3.6}$$

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, refer to Section 4.2 for a detailed discussion of the method. 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 $b = 4$.

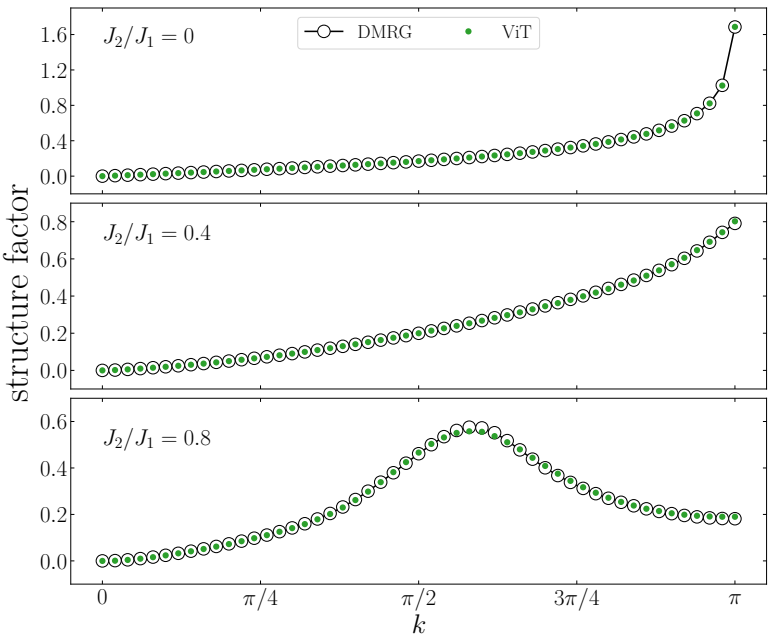


Figure 3.4: The spin-spin structure factor $S(k)$ as computed by the ViT wave function (full dots) for $J_2/J_1 = 0$ (upper panel), $J_2/J_1 = 0.4$ (middle panel) and $J_2/J_1 = 0.8$ (lower panel) on a cluster with $L = 100$ sites. The DMRG results are also shown for comparison (empty circles).

3.3 ViT for the J_1-J_2 Heisenberg chain

We start by discussing how the accuracy of the ViT wave function with one layer can be systematically improved by varying its two hyper-parameters, 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 [124]). In Fig. 3.2, 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 [122, 126–129], 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 $J_2/J_1 = 0.4$.

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

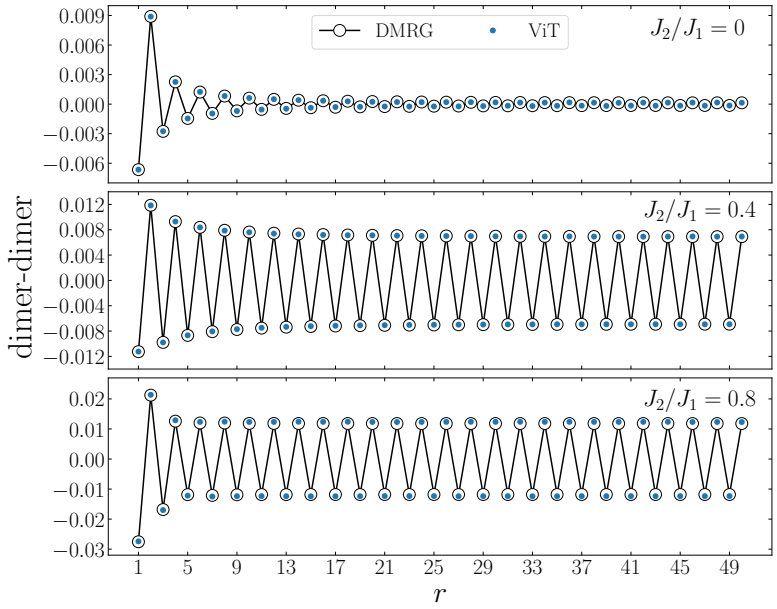


Figure 3.5: Dimer-dimer correlations as computed by the ViT wave function (full circles) for $J_2/J_1 = 0$ (upper panel), $J_2/J_1 = 0.4$ (middle panel) and $J_2/J_1 = 0.8$ (lower panel) on a cluster with $L = 100$ sites. The DMRG results are also shown for comparison (empty circles).

and complexity, for which the network can be trained on $L = 100$ sites in a few hours on ten CPUs or in a few minutes on a GPU. The spin-spin correlations are defined as

$$C^{\nu\nu}(r) = \frac{1}{L}\sum_{R=0}^{L-1}\langle \hat{S}^{\nu}_{R}\hat{S}^{\nu}_{R+r}\rangle, \tag{3.7}$$

where $\nu = x$, y, or z and $\langle \dots \rangle$ 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 $S(k) = \frac{1}{L}\sum_{r=0}^{L-1} \mathrm{e}^{ikr} C(r)$. In Fig. 3.3, we show the results of the real-space correlations $C(r)$ for the 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 multi-head 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. 3.4.

The gapped phase is characterized by a finite dimer order (implied by the two-

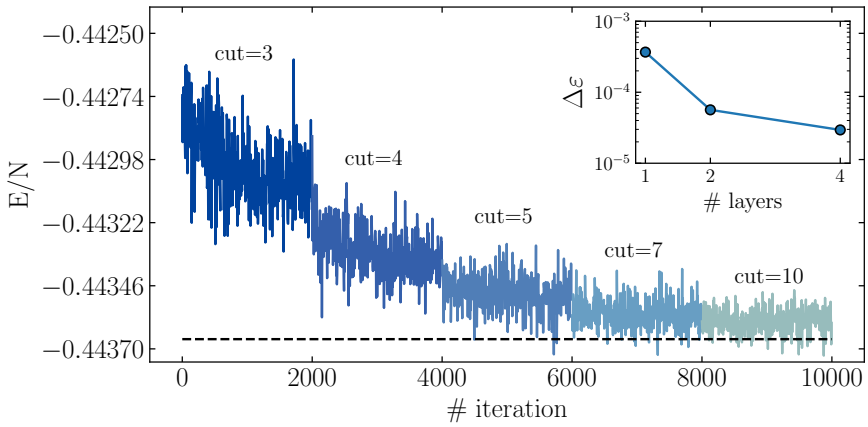


Figure 3.6: Optimization of the DeepViT with $n_l = 4$, where each layer has $h = 2$ and $d/h = 2$, for the Heisenberg model with $L = 40$ sites. Along the process, the cut in the attention is fixed and progressively increased from 1 to 10 (the first two values are not shown for better readability). At the end, once the cut has been completely relaxed, the full translational invariance is restored [see Eq. (3.6)] 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 [124] with a bond dimension up to $\chi = 600$ obtaining $E/J_1 = -0.443663$.

fold 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:

$$D(r) = \frac{1}{L}\sum_{R=0}^{L-1}\langle \hat{S}^z_R \hat{S}^z_{R+1} \hat{S}^z_{R+r} \hat{S}^z_{R+r+1}\rangle - [C^{zz}(r=1)]^2, \tag{3.8}$$

where $C^{zz}(r = 1)$ is the z component of the spin-spin correlation function at distance $r = 1$ defined in eq. (3.7). Notice that this definition considers only the z component of the spin operators [130]. In Fig. 3.5, 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.

3.4 Relax-cut procedure to optimize deep complex-valued ViT

The ViT wave function can be systematically improved by stacking multiple Transformer layers, i.e., $n_l > 1$. The optimization of deep networks employing complex valued parameters is difficult with standard protocols, then we develop the "relax-cut" procedure based on the physical interpretation of the attention weights. We start by setting for each head and layer $\alpha_{i-j} = 0$ if $|i - j| >$ cut, with 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. 3.6. 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.

3.5 Discussion

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*, are 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 of only one layer, excellent results are obtained for a frustrated spin model in one spatial dimension.

Chapter 4

Feature learning perspective on Neural Quantum States

Understanding quantum magnetism in two-dimensional systems represents a lively branch in modern condensed-matter physics. In the presence of competing super-exchange couplings, magnetic order is frustrated and can be suppressed down to zero temperature. Still, capturing the correct nature of the exact ground state is a highly complicated task, since energy gaps in the spectrum may be very small and states with different physical properties may have competing energies. In this Chapter, we first introduce in Section 4.1 a variational *Ansatz* for two-dimensional frustrated magnets by leveraging the power of representation learning. A central innovation of our approach is to overcome the limitations of the *Ansatz* introduced in the previous chapter, where stable optimization of deep networks was hindered by the presence of complex-valued parameters (see Section 3.4). To address this, we employ a deep Vision Transformer with real-valued parameters to map physical spin configurations into a high-dimensional feature space. Within this abstract space, the determination of the ground-state properties is simplified and requires only a shallow output layer with complex-valued parameters.

Crucially, the proposed ViT *Ansatz* contains a large number of parameters due to its deep architecture, which makes standard optimization methods impractical. To address this, we develop a scalable and efficient formulation of the Stochastic Reconfiguration method in Section 4.2. Using this novel optimization protocol together with our *Ansatz*, in Section 4.3 we illustrate its efficacy by studying the J_1-J_2 Heisenberg model at $J_2/J_1 = 0.5$ on the 10×10 square lattice, a challenging benchmark in highly-frustrated magnetism, and achieving state-of-the-art ground-state energy by optimizing a ViT architecture with 3×10^5 parameters.

4.1 Feature learning inspired Ansatz

In this Section, we take a new perspective on NQS by leveraging the principle of *representation learning* [131] that is key to the success of deep neural networks in practice. The idea is that the mathematical structure of deep networks, a composition of simple functions with parameters that can be tuned to data, allows neural networks to automatically extract the pertinent features of a data set for a given task. These features or *representations* of the inputs are then used for downstream tasks, like, in our case, predicting the amplitude of the wave function for a given spin configuration. The idea of *learning* these representations directly from data is contrary to the approach of classical machine learning, which required careful engineering and considerable domain expertise to distil raw data (such as the spin configurations) into a representation or feature vector that could be used for a downstream task [132].

In this Section, we follow this approach by building a variational *Ansatz* where we use a deep neural network to map physical spin configurations into a feature space. This transformation enables an accurate prediction of the amplitude associated with each configuration with even a simple, shallow fully-connected layer [17]. By reframing the NQS as feature extractors rather than just universal approximators of complicated functions, the variational state is naturally perceived as the composition of two distinct functions, each with a specific role:

$$\begin{aligned} \boldsymbol{z} &= V(\boldsymbol{\sigma};\phi) \ , \\ \mathrm{Log}[\Psi_\theta(\boldsymbol{\sigma})] &= f\left(\boldsymbol{z};W\right) \ , \end{aligned} \tag{4.1}$$

where the variational parameters are partitioned into two blocks $\theta = \{\phi, W\}$. The function $V(\cdot;\phi)$ is parameterized as a *deep* neural network, mapping physical configurations $\boldsymbol{\sigma}$ to vectors $\boldsymbol{z}$, called *hidden representations*, which belong to a d-dimensional *feature space*. Conversely, $f(\cdot;W)$ is a *shallow* neural network used to generate a single scalar value from the hidden representations $\boldsymbol{z}$. This final value is used to predict the amplitude corresponding to the input configuration. In order to predict both modulus and phase of the variational state (which is fundamental in cases where the exact sign is not known *a priori*), it is convenient to employ a complex-valued variational state. The structure of the *Ansatz* in Eq. (4.1) suggests the possibility of taking ϕ as real-valued parameters in the deep neural network $V(\cdot;\phi)$. Subsequently, only the parameters W of the shallow function $f(\cdot;W)$ can be taken complex-valued. We schematically represent these two steps in the left panel of Fig. 4.1; instead, a pictorial scheme of the mapping process from the physical space of the spin configurations to the feature space is depicted in the right panel of Fig. 4.1.

Far from being only a change of viewpoint, the possibility of having a real-valued feature extractor is crucial in practice. Several works showed recently

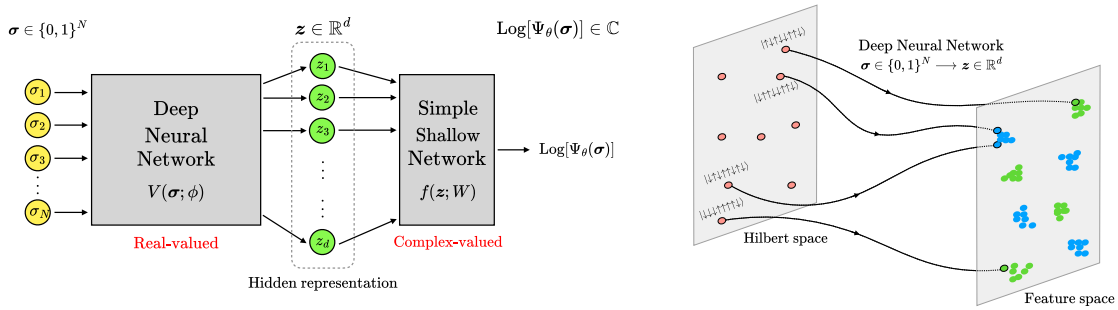


Figure 4.1: **Left panel:** The NQS is defined as the composition of two functions: first, a deep neural network $V(\boldsymbol{\sigma};\phi)$ (with real-valued parameters) maps the input configurations $\boldsymbol{\sigma}$ into hidden representations $\boldsymbol{z}$; then, a simple shallow network $f(\boldsymbol{z};W)$ (with complex-valued parameters) generates the logarithm of the amplitudes $\mathrm{Log}[\Psi_\theta(\boldsymbol{\sigma})]$ starting from hidden representations. **Right panel:** Pictorial illustration of the mapping process carried out by the deep neural network. The network maps spin configurations from the Hilbert space $\boldsymbol{\sigma}$ into configurations in a feature space $\boldsymbol{z} \in \mathbb{R}^d$ with a non-trivial structure. In subsection 4.1.3, we show for example that physical configurations $\boldsymbol{\sigma}$ cluster in feature space according to the sign of the amplitudes $\mathrm{Log}[\Psi_\theta(\boldsymbol{\sigma})]$.

that depth is crucial to achieving high accuracies on two-dimensional quantum systems [22, 23, 133, 134]. Yet, training deep networks is a complicated task that is only possible by leveraging techniques such as Layer Normalization [135], skip connections [136], and appropriate activation functions [137]. However, all of these techniques have been developed for real-valued architectures, and cannot be straightforwardly generalized to complex-valued neural networks. For these reasons, in Section 3.4, the optimization of a deep Transformer architecture having complex-valued parameters necessitated the development of a heuristic procedure involving the introduction of a cut in the attention weights. A big advantage of the newly proposed *Ansatz* with a real-valued feature extractor is then that it can be trained from scratch without additional restrictions and with minimal regularization in the optimization protocol (Section 4.2 for details). This modified architecture has recently yielded state-of-the-art results on one of the most popular benchmarks in frustrated magnetism [37]. The following two subsections give a detailed description of the architecture of the neural network.

4.1.1 Vision Transformer

We propose the use of a Vision Transformer [105] (see Fig. 4.2) to parametrize $V(\cdot;\phi)$ in Eq. (4.1), see Section 4.1.2 for a detailed description of the architecture;

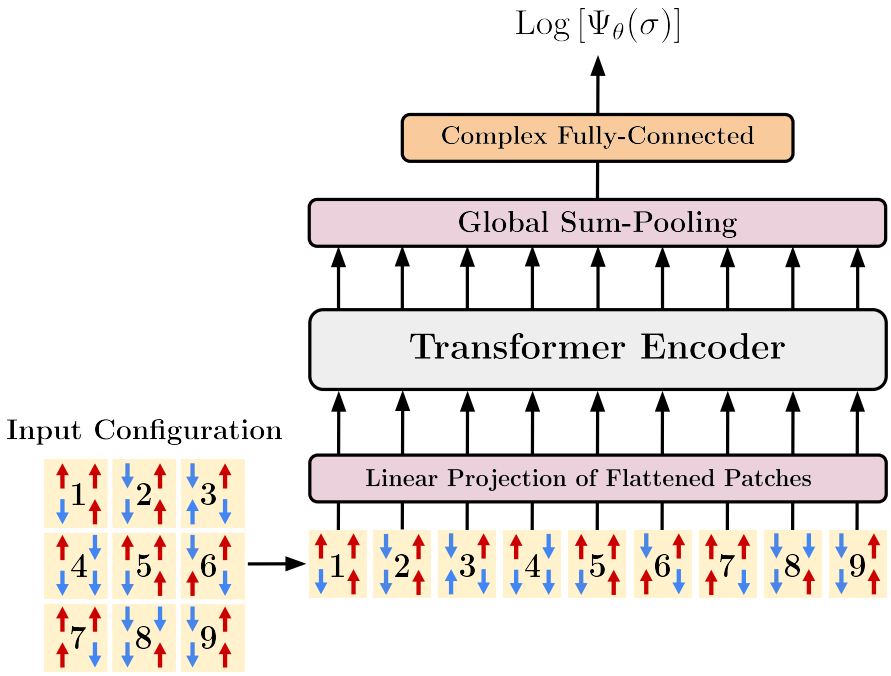


Figure 4.2: The input spin configuration $\boldsymbol{\sigma}$ is partitioned into patches, which are linearly projected in a d-dimensional embedding space and then processed by a Vision Transformer. The latter one builds new representations of the patches, which are then combined through summation and fed into a final single complex-valued fully-connected layer in order to obtain the logarithm of the (complex) wave function. Notice that this is a particular instantiation of the more general scheme proposed in the left panel of Fig. 4.1.

instead, the function f is chosen to be:

$$f(\boldsymbol{z}; W) = \sum_{\alpha=1}^{K} \log\cosh\left(b_\alpha + \boldsymbol{w}_\alpha \cdot \boldsymbol{z}\right) \ , \tag{4.2}$$

where the variational parameters W are the biases and the weights of the linear transformation. The number of hidden neurons K is a hyperparameter of the network. Notice that Eq. (4.2) has the same functional form as the well-known Restricted-Boltzmann Machine (RBM) introduced by Carleo and Troyer [17]. Crucially, in this case it is not applied to the physical configuration $\boldsymbol{\sigma}$ but instead to the hidden representation $\boldsymbol{z}$. This is the change of paradigm that we want to emphasize. With these choices, the process of constructing the amplitude corresponding to a physical spin configuration $\boldsymbol{\sigma}$ involves the following steps (see Fig. 4.2):

1. The input spin configuration $\boldsymbol{\sigma}$ is initially divided into n patches (see Ap-

pendix 8.4.1).

2. The patches are linearly projected into a d-dimensional embedding space, resulting in a sequence of vectors $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$.

3. A ViT processes these embedded patches, producing another sequence of vectors $(\mathbf{y}_1, \cdots, \mathbf{y}_n)$, where $\mathbf{y}_i \in \mathbb{R}^d$.

4. The hidden representation $\boldsymbol{z}$ of the configuration $\boldsymbol{\sigma}$ is defined by summing all these output vectors: $\boldsymbol{z} = \sum_{i=1}^{n} \mathbf{y}_i$.

5. A fully-connected layer with complex-valued parameters, defined in Eq. (4.2), produces the amplitude $\mathrm{Log}[\Psi_\theta(\boldsymbol{\sigma})]$ corresponding to the input configuration $\boldsymbol{\sigma}$. Specifically, we set $K = d$.

Notably, while the vector $\mathbf{x}_i$ depends solely on the spins contained in the i-th patch, the resulting vector $\mathbf{y}_i$ is a function of all the spins in the configuration. The ViT architecture is constructed as a sequence of n_l encoder blocks. In each of them, a multi-head factored attention layer (with h heads) is followed by a two-layer fully connected network. For a detailed description of the Encoder Block see subsection 4.1.2.

Notice that the structure of this variational *Ansatz* requires a large number of parameters. In order to optimize them, in Section 4.2 we introduce a modern formulation of the Stochastic Reconfiguration technique [138] to be able to deal with a large number of variational parameters [22, 37].

4.1.2 Encoder Block

The Encoder Block is the core of the Transformer architecture (see Fig. 4.3). The input sequence of the l-th Encoder Block (where l runs from 1 to n_l) is the set of n vectors $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where, for the sake of simplicity, the index l is not made explicit. This sequence of vectors is processed by a real-valued factored multi-head attention mechanism [37, 139], with h the number of heads in the multi-head attention mechanism. The μ-th attention vector $\mathbf{A}_i^\mu$ is defined by first applying a *local* linear transformation $\boldsymbol{V}^\mu$ to each input vector $\boldsymbol{x}_j$. The resulting vectors $V^\mu \mathbf{x}_j$ are then *globally* mixed according to the factored attention mechanism introduced in Eq. (3.5). The crucial difference with respect to the previous ViT parameterization is that the matrices of parameters α (which define the so-called *attention maps*) and V are real valued instead of complex valued. The h different attention representations computed in each head $\mathbf{A}_i^\mu \in \mathbb{R}^{d/h}$ are concatenated together to give an output sequence of n attention vectors $(\mathbf{A}_1, \cdots, \mathbf{A}_n)$, with $\mathbf{A}_i \in \mathbb{R}^d$. Then, after another linear projection which mixes the representations

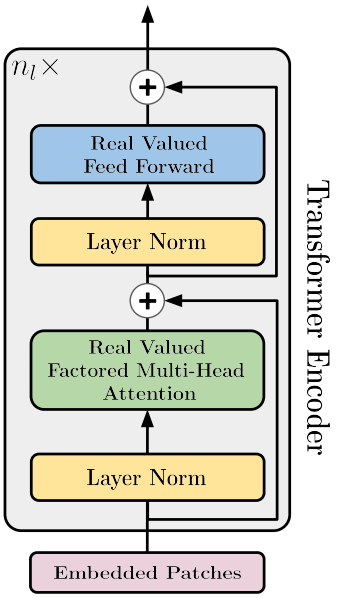


Figure 4.3: To process the embedded patches, each Transformer Encoder block employs a real-valued factored multi-head attention mechanism, which mixes the patches, and a real-valued two-layers Feed-Forward neural network, which is used to introduce a non-linearity. Skip connections and Layer Normalization are also employed.

of the different heads, each attention vector is finally passed identically and independently through a non-linearity, which is taken to be a (real-valued) two-layers fully-connected neural network, with hidden dimension $2d$ and the standard rectified linear unit (ReLU) activation function. The output of the l-th encoder block is a sequence $(\mathbf{y}_1, \cdots, \mathbf{y}_n)$, with $\mathbf{y}_i \in \mathbb{R}^d$ being a new representation of the i-th input. Pre-Layer Normalization [135] and skip connections are used, these being the key elements that permit the optimization of deep networks. The use of factored attention in Eq. (3.5) is justified by the physical interpretation we give to the attention weights; indeed, we expect that they should mainly depend on the relative positions among groups of spins and not on the actual values of the spins in the patches [21, 139]. Moreover, attention weights are taken translationally invariant, in order to encode the translational symmetry between patches.

4.1.3 Hidden representations

The motivation for the ViT wave function introduced in this Section is to leverage the power of representation learning. Instead of using a neural network simply as a universal approximator to map spin configurations to wave function amplitudes,

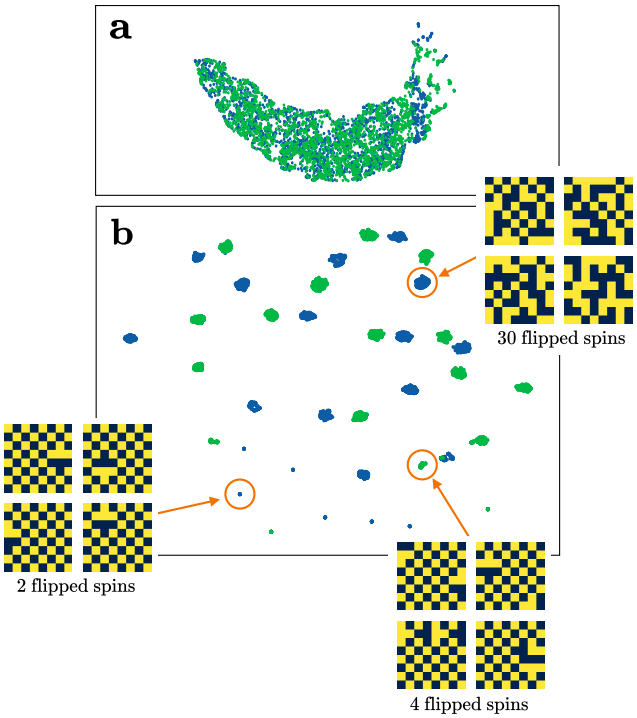


Figure 4.4: Dimensional reduction (using the UMAP algorithm) of the hidden representations for a set of configurations obtained from a ViT in the limit of $J' = 0$, leading to the Heisenberg model. Points are colored according to the exact signs given by the Marshall sign rule [120]. a) Before optimizing, when the parameters of the neural network are random, the hidden representations are all concentrated in a single cluster. b) After the optimization of the variational energy the hidden representations are organized in clusters with the same number of flipped spins with respect to the Néel state. All calculations are performed on an 8×8 cluster.

we train the network to map spin states into a feature space. These features are then used as an input to a shallow neural network to predict the amplitude corresponding to the input configuration [see Eq. (4.2)]. This change in perspective has important practical ramifications: it dispenses with the need of having complex-valued parameters in the feature extractor $V(\boldsymbol{\sigma}; \phi)$, making the training of deep networks much easier. In this section, we verify that the network indeed learns a set of non-trivial features in the course of training, and we find that some of these features are even interpretable in a simple limiting case.

To this end, we examine the Heisenberg model on a two dimensional square lattice; here, the ground state properties are characterized by the presence of antiferromagnetic order. In addition, the exact sign structure of the ground state

follows the Marshall sign rule [120]. For a given set of configurations $\{\boldsymbol{\sigma}_i\}$ (sampled along the Monte Carlo procedure), we compute the corresponding hidden vectors $\{\mathbf{z}_i\}$ of size $d \gg 1$, which can be visualized in two dimensions after a dimensional reduction with the standard Uniform Manifold Approximation and Projection method (UMAP) [140].

In Fig. 4.4, we color each feature or representation $\mathbf{z}_i$ according to the exact sign of the amplitude corresponding to the spin configuration $\boldsymbol{\sigma}_i$. Before the Transformer is trained, i.e., with random network weights, there is no discernible structure in the representations (see Fig. 4.4a). After training, by means of minimizing the variational energy, we see instead that features have a highly non-trivial structure: they are grouped into different clusters of the representation space (see Fig. 4.4b). This is a direct result of the training.

Remarkably, these clusters have a physical interpretation in the unfrustrated case, where the spin configurations in a given cluster have the same number of flipped spins with respect to the Néel state and, therefore, the same sign (according to the Marshall rule), and similar modulus. The crucial point is that, by using a single fully-connected layer, the prediction of the correct amplitudes is much easier when acting on these representations rather than using the original spin configurations, as can be seen from the vastly superior energies obtained by the ViT state compared to a simple RBM [37, 141].

A similar clustering structure in the trained ViT also appears for a generic Hamiltonian, though the precise interpretation of the clusters is less straightforward (see Appendix 8.4.4. This reflects the situation in machine learning, where it is generally difficult to extract human-interpretable structure directly from the representations. In fact, the whole point of letting neural networks learn features directly from data is that they can discover subtle patterns in the data that are hard for humans to extract or even describe, but nevertheless turn out to be useful for the task at hand.

4.2 A simple linear algebra identity to optimize Large-Scale Neural Network Quantum States

The network developed in Section 4.1 requires a large number of variational parameters. However, the Stochastic Reconfiguration (SR) method [16, 108, 109] (see below), which is essential to obtain stable and accurate optimizations of Neural Quantum States, becomes impractical beyond a few thousand parameters. In this Section, we leverage a simple linear algebra identity to show that SR can be employed even in the deep learning scenario.

The Eq. (3.2) defines the standard formulation of the SR, which involves the

inversion of a $P \times P$ matrix, being the bottleneck of this approach when the number of parameters is larger than $O(10^4)$. To address this problem, we start reformulating Eq. (3.2) in a more convenient way. For a given sample of M spin configurations $\{\sigma_i\}$ (sampled according to $|\Psi_\theta(\sigma)|^2$), the stochastic estimate of F_α [see Eq. (3.1)] can be obtained as:

$$\bar{F}_\alpha = -\,\mathrm{Re}\left[\frac{2}{M}\sum_{i=1}^{M}[E_{Li} - \bar{E}_L]^*[O_{\alpha i} - \bar{O}_\alpha]\right] \ . \tag{4.3}$$

Here, $E_{Li} = \langle\sigma_i|\hat{H}|\Psi_\theta\rangle \,/\, \langle\sigma_i|\Psi_\theta\rangle$ defines the local energy for the configuration $|\sigma_i\rangle$ and $O_{\alpha i} = O_\alpha(\sigma_i)$; in addition, $\bar{E}_L$ and $\bar{O}_\alpha$ denote sample means. Throughout this work, we adopt the convention of using latin and greek indices to run over configurations and parameters, respectively. Equivalently, Eq. (3.3) can be stochastically estimated as

$$\bar{S}_{\alpha,\beta} = \mathrm{Re}\left[\frac{1}{M}\sum_{i=1}^{M}\left[O_{\alpha i} - \bar{O}_\alpha\right]^*\left[O_{\beta i} - \bar{O}_\beta\right]\right] \ . \tag{4.4}$$

To simplify further, we introduce $Y_{\alpha i} = (O_{\alpha i} - \bar{O}_\alpha)/\sqrt{M}$ and $\varepsilon_i = -2[E_{Li} - \bar{E}_L]^*/\sqrt{M}$, allowing us to express Eq. (4.3) in matrix notation as $\bar{\boldsymbol{F}} = \mathrm{Re}[Y\boldsymbol{\varepsilon}]$ and Eq. (3.3) as $\bar{S} = \mathrm{Re}[YY^\dagger]$. Writing $Y = Y_R + iY_I$ we obtain:

$$\bar{S} = Y_R Y_R^T + Y_I Y_I^T = XX^T \tag{4.5}$$

where $X = \mathrm{Concat}(Y_R, Y_I) \in \mathbb{R}^{P\times 2M}$, the concatenation being along the last axis. Furthermore, using $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_R + i\boldsymbol{\varepsilon}_I$, the gradient of the energy can be recast as

$$\bar{\boldsymbol{F}} = Y_R\boldsymbol{\varepsilon}_R - Y_I\boldsymbol{\varepsilon}_I = X\boldsymbol{f} \ , \tag{4.6}$$

with $\boldsymbol{f} = \mathrm{Concat}(\boldsymbol{\varepsilon}_R, -\boldsymbol{\varepsilon}_I) \in \mathbb{R}^{2M}$. Then, the update of the parameters in Eq. (3.2) can be written as

$$\boldsymbol{\delta\theta} = \tau(XX^T + \lambda\mathbb{I}_P)^{-1}X\boldsymbol{f} \ . \tag{4.7}$$

This reformulation of the SR updates is a crucial step, which allows the use of the well-known push-through identity $(AB + \lambda\mathbb{I}_n)^{-1}A = A(BA + \lambda\mathbb{I}_m)^{-1}$ [142, 143], where A and B are respectively matrices with dimensions $n \times m$ and $m \times n$. As a result, Eq. (4.7) can be rewritten as

$$\boldsymbol{\delta\theta} = \tau X(X^TX + \lambda\mathbb{I}_{2M})^{-1}\boldsymbol{f} \ . \tag{4.8}$$

This derivation shows, in a simple and transparent way, how to exactly perform the SR with the inversion of a $2M \times 2M$ matrix and, therefore, without allocating a $P \times P$ matrix. We emphasize that the last formulation is very useful in the typical deep learning setup, where $P \gg M$. Employing Eq. (4.8) instead

of Eq. (4.7) proves to be more efficient in terms of both computational complexity and memory usage. The required operations for this new formulation are $O(M^2P) + O(M^3)$ instead of $O(P^3)$, and the memory usage is only $O(MP)$ instead of $O(P^2)$. For deep neural networks with n_l layers the memory usage can be further reduced roughly to $O(MP/n_l)$ (see Ref. [144]). Other methods, based on iterative solvers, require $O(nMP)$ operations, where n is the number of steps needed to solve the linear problem in Eq. (3.2). However, this number increases significantly for ill-conditioned matrices (the matrix S has a number of zero eigenvalues equal to $P - M$), leading to many non-parallelizable iteration steps and consequently higher computational costs [145]. Our proof also highlights that the diagonal-shift regularization of the S matrix in parameter space [see Eq. (3.2)] is equivalent to the same diagonal shift in sample space [see Eq. (4.8)]. A similar result has been obtained in Ref [146], with the difference that a pseudo-inverse regularization is applied in order to truncate the effect of vanishing singular values during inversion.

4.3 State of the art results on the J_1-J_2 Heisenberg model

In this Section, we combine the Deep ViT variational *Ansatz* introduced in Section 4.1 and the optimization method developed in Section 4.2 to address the most popular benchmark problem in the field of frustrated quantum many-body spin systems: the approximation of the ground state of the J_1-J_2 Heisenberg model in the highly frustrated point $J_2/J_1 = 0.5$ on the 10×10 square lattice. The result in Table 4.1 is achieved with the Deep ViT using $b = 2$, $n_l = 8$ layers, embedding dimension $d = 72$, and $h = 12$ heads per layer. This variational state has in total 267720 real parameters (the complex-valued parameters of the output layer are treated as couples of independent real-valued parameters). Regarding the optimization protocol, we choose the learning rate $\tau = 0.03$ (with cosine decay annealing) and the number of samples is fixed to be $M = 6000$. We emphasize that using Eq. (4.7) to optimize this number of parameters would be infeasible on available GPUs: the memory requirement would be more than $O(10^3)$ gigabytes, one order of magnitude bigger than the highest available memory capacity. Instead, with the formulation of Eq. (4.8), the memory requirement can be easily handled by available GPUs. The simulations took four days on twenty A100 GPUs. Remarkably, as illustrated in Table 4.1, we are able to obtain the state-of-the-art ground-state energy using an architecture solely based on neural networks, without using any other regularization than the diagonal shift reported in Eq. (4.8), fixed to $\lambda = 10^{-4}$. We stress that a completely unbiased simulation, without as-

Table 4.1: Ground-state energy on the 10×10 square lattice at $J_2/J_1 = 0.5$.

Energy per site	**Wave function**	**# parameters**	**Marshall prior**	**Reference**	**Year**
-0.48941(1)	MLP	893994	Not available	[147]	2023
-0.494757(12)	CNN	Not available	No	[148]	2020
-0.4947359(1)	Shallow CNN	11009	Not available	[149]	2018
-0.49516(1)	Deep CNN	7676	Yes	[150]	2019
-0.495502(1)	PEPS + Deep CNN	3531	No	[151]	2021
-0.495530	DMRG	8192 SU(2) states	No	[152]	2014
-0.495627(6)	aCNN	6538	Yes	[153]	2023
-0.49575(3)	RBM-fermionic	2000	Yes	[154]	2019
-0.49586(4)	CNN	10952	Yes	[155]	2023
-0.4968(4)	RBM ($p = 1$)	Not available	Yes	[156]	2022
-0.49717(1)	Deep CNN	106529	Yes	[157]	2022
-0.497437(7)	GCNN	67548	No	[158]	2023
-0.497468(1)	Deep CNN	421953	Yes	[159]	2022
-0.4975490(2)	VMC ($p = 2$)	5	Yes	[160]	2013
-0.497627(1)	Deep CNN	146320	Yes	[146]	2023
-0.497629(1)	RBM+PP	13132	Yes	[161]	2021
-0.497634(1)	**Deep ViT**	**267720**	**No**	**Ours**	**2023**

suming any prior for the sign structure, is performed, in contrast to other cases where the Marshall sign rule is used to stabilize the optimization [146, 150, 157, 159, 161] (see Table 4.1). Furthermore, we verified with numerical simulations that the final results is not affected by the Marshall prior. This is an important point since a simple sign prior is not available for the majority of the models (e.g., the Heisenberg model on the triangular or Kagome lattices). We would like also to mention that the definition of a suitable architecture is fundamental to take advantage of having a large number of parameters. Indeed, while a stable simulation with a simple regularization scheme (only defined by a finite value of λ) is possible within the Deep ViT wave function, other architectures require more sophisticated regularizations. For example, to optimize Deep GCNNs it is necessary to add a temperature-dependent term to the loss function [158] or, for Deep

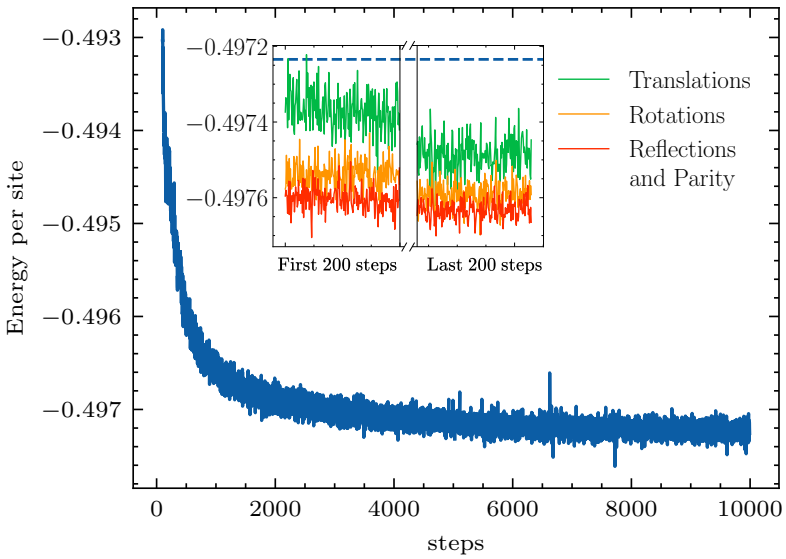


Figure 4.5: **Variational energy optimization.** Optimization of the Deep ViT with patch size $b = 2$, $n_l = 8$ layers, embedding dimension $d = 72$ and $h = 12$ heads per layer, on the J_1-J_2 Heisenberg model at $J_2/J_1 = 0.5$ on the 10×10 square lattice. The first 200 optimization steps are not shown for better readability. Inset: first and last 200 optimization steps when recovering sequentially the full translational (green curve), rotational (orange curve) and reflections and parity (red curve) symmetries. The total number of steps after restoring the symmetries is 5000 for translations, 5000 for rotations and 4000 for reflections and parity. The mean energy obtained without quantum number projection is also reported for comparison (blue dashed line).

CNNs, a process of variance reduction and reweighting [146] helps in escaping local minima. We also point out that physically inspired wave functions, as the Gutzwiller-projected states [160], which give a remarkable result with only a few parameters, are not always equally accurate in other cases.

In Fig. 4.5 we show a typical Deep ViT optimization on the 10×10 lattice at $J_2/J_1 = 0.5$. First, we optimize the Transformer having translational invariance among patches (blue curve). Then, starting from the previously optimized parameters, we restore sequentially the full translational invariance (green curve), rotational symmetry (orange curve) and lastly, reflections and spin parity symmetry (red curve). The symmetrized wave function reads:

$$\tilde{\Psi}_\theta(\sigma) = \sum_{i=0}^{b^2-1} \sum_{j=0}^{7} \left[\Psi_\theta(T_i R_j \sigma) + \Psi_\theta(-T_i R_j \sigma)\right] \ . \tag{4.9}$$

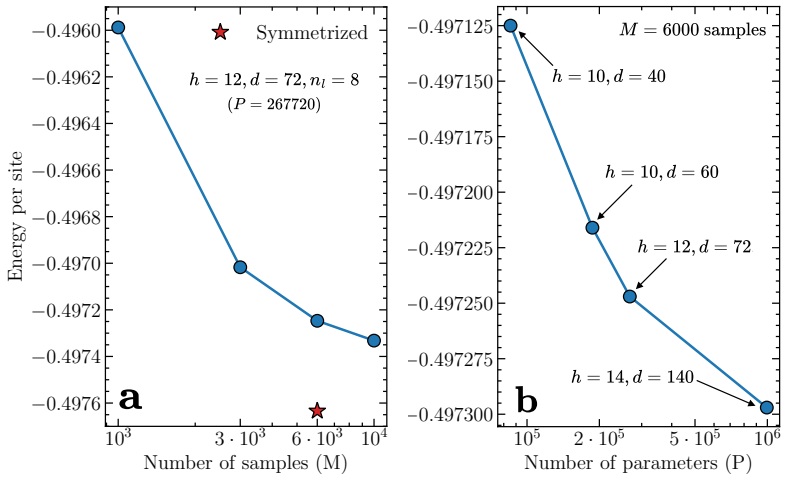


Figure 4.6: **Variational energy scalings.** Panel **a**: Energy per site as a function of the number of samples M for a ViT with $n_l = 8$ layers, embedding dimension $d = 72$ and $h = 12$ heads per layer. Panel **b**: Energy per site as a function of the number of parameters P, increased by adding heads h and taking larger embedding dimensions d, for a fixed number of layers $n_l = 8$. For both panels, the energy values (blue circles) are obtained without restoring the symmetries; for comparison we also show the energy corresponding to the fully symmetrized state in Eq. (4.9) (red star) which is the one reported in Table 4.1.

In the last equation, T_i and R_j are the translation and the rotation/reflection operators, respectively. Whenever a new symmetry is restored, the energy reliably decreases [162]. We stress that our optimization process, which combines the SR formulation of Eq. (4.8) with a real-valued Deep ViT followed by a complex-valued fully connected output layer [21], is highly stable and insensitive to the initial seed, ensuring consistent results across multiple optimization runs.

Finally, we discuss the impact of the number of samples M and parameters P on the variational energy of the ViT wave function, showing the results in Fig. 4.6. In the right panel, we show the variational energy as a function of the number of parameters for a fixed number of layers $n_l = 8$, performing the optimizations with $M = 6000$ samples. The number of parameters is increased by enlarging the width of each layer. In particular, we take the following architectures: $(h = 10, d = 40)$, $(h = 10, d = 60)$, $(h = 12, d = 72)$, and $(h = 14, d = 140)$ with $P = 85400$, 187100, 267720, and 994700 parameters, respectively. A similar analysis for a fixed width but varying the number of layers is discussed in Ref. [21]. Instead, in the left panel, we fix an architecture $(h = 12, d = 72)$ with $n_l = 8$ and increase the number of samples M up to 10^4. Both analyses are performed without restoring the symmetries by quantum number projection; for comparison, we report in the

left panel the energy obtained after restoring the symmetries [see Eq. (4.9)]. The latter one coincides with the ViT wave function used to obtain the energy reported in Table 4.1. We point out that the energy curves depicted in Fig. 4.6 are obtained from unbiased simulations, without the utilization of the Marshall sign prior. The final value of the energy, as well as the convergence to it, are qualitatively similar regardless of its inclusion.

4.4 Discussion

In this chapter, with calculations on the most popular benchmark in the field of many-body quantum systems, we demonstrate that the ViT *Ansatz* rises among the universe of variational wave functions as a possible way to eventually solve important quantum many-body problems. In order to optimize deep architectures with a large number of parameters, we have introduced a formulation of the Stochastic Reconfiguration method that excels in scenarios where the number of parameters (P) significantly outweighs the number of samples (M) used for stochastic estimations. One key feature is the ability of the ViT wave function to create a mapping of the physical configurations in a *real* feature space, where it is then easy to predict amplitudes, even with a single fully-connected layer. Looking at NQSs as feature extractors is the most important original contribution of this chapter, in contrast with the common interpretation of just universal approximators of functions, which usually leads to taking all the parameters complex-valued.

Chapter 5

The Shastry-Sutherland model

In this Chapter, we use the ViT architecture proposed in Chapter 4 to study the ground-state phase diagram of the Shastry-Sutherland model, which captures the low-temperature behavior of $SrCu_2(BO_3)_2$ with its intriguing properties. With highly accurate numerical simulations, we provide strong evidence for the stabilization of a spin-liquid between the plaquette and antiferromagnetic phases. In addition, a direct calculation of the triplet excitation at the Γ point provides compelling evidence for a gapless spin liquid. Remarkably, the characterization of the spin liquid phase within the Shastry-Sutherland model has not been explored in previous studies. Our results are particularly important because they show that the magnetically ordered Néel phase is melted into a gapless spin liquid, similarly to what happens in the J_1-J_2 Heisenberg model on the square lattice [41]. This suggests that this kind of (continuous) transition is rather generic and may represent the habit, and not the exception, for the melting of the Néel order due to magnetic frustration. In addition, our calculations clearly demonstrate that the ViT *Ansatz* rises among the universe of variational wave functions as a possible way to eventually solve important quantum many-body problems, and open up new possibilities for establishing the properties of many-body systems.

5.1 Emergence of a gapless spin liquid phase in the Shastry-Sutherland model

Since the discovery of the fractional quantum Hall effect [163] and its description by the Laughlin wave function [107], a growing interest has developed around unconventional phases of matter, i.e., the ones that escape perturbative or mean-field approaches. In this sense, the hunt for spin liquids is of fundamental importance in Mott insulators, where localized spins determine the low-temperature properties. On geometrically frustrated lattices, it is not possible to minimize simultaneously

all the interactions among the spins and, therefore, magnetic order could be suppressed, even at zero temperature. In this case, spins are highly entangled and the resulting ground-state wave function shows unconventional properties [164]. However, most of the theoretical models that have been proposed to support quantum spin liquids are still unresolved, and their phase diagrams are not well established except for specific points (that usually give trivial states). One notable exception is given by the Kitaev model on the honeycomb lattice [165], which provides a formidable example for gapless and gapped spin liquids. On the experimental side, there has been great development in the search for materials that might be able to support these exotic phases of matter. One promising example is given by the so-called Herbertsmithite, which may realize a spin liquid at low temperatures [166]. Among the variety of quantum spin models, the one introduced by Shastry and Sutherland [38] deserves particular attention since it gives an example in which the magnetic order can be melted by tuning the super-exchange interactions, leading to a particularly simple ground-state wave function, where nearby spins form singlets. Most importantly, this Hamiltonian captures the low-temperature properties of $SrCu_2(BO_3)_2$ [39, 40].

The main interest in this material comes from its properties when external magnetic fields are applied. Indeed, a complicated magnetization curve is observed, with various magnetization plateaus (most notably at magnetization 1/8) that show intriguing properties [39, 167–169]. The Shastry-Sutherland model is defined by

$$\hat{H} = J \sum_{\langle \mathbf{r},\mathbf{r}' \rangle} \hat{\mathbf{S}}_{\mathbf{r}} \cdot \hat{\mathbf{S}}_{\mathbf{r}'} + J' \sum_{\langle\langle \mathbf{r},\mathbf{r}' \rangle\rangle} \hat{\mathbf{S}}_{\mathbf{r}} \cdot \hat{\mathbf{S}}_{\mathbf{r}'} \,, \tag{5.1}$$

where $\hat{\mathbf{S}}_{\mathbf{r}}$ is the $S = 1/2$ operator on the site $\mathbf{r}$ of a $L \times L$ square lattice, with periodic boundary conditions. Here, the first sum goes over nearest-neighbor sites on the square lattice, while the second sum is over next-nearest-neighbor sites on orthogonal dimers, according to the bond pattern of Fig. 5.1. For a detailed description of the lattice structure, including its symmetries, see Appendix 8.4.1.

The ground-state properties of the Shastry-Sutherland model are well known in two limiting cases. When $J = 0$, the model reduces to a collection of decoupled dimers and its ground state is a product of singlets connected by J'; this state remains the exact ground state also for finite values of J/J', up to a certain value [38]. In the opposite limit, when $J' = 0$, the Heisenberg model on the square lattice is recovered, whose ground state is the Néel antiferromagnet; also in this case, the ground state is robust in a finite region when $J' > 0$. Despite the substantial effort that has been invested in understanding the appearance of magnetization plateaus, the ground-state properties of the Shastry-Sutherland model have been investigated in much less depth. One of the first studies based on the mean-field

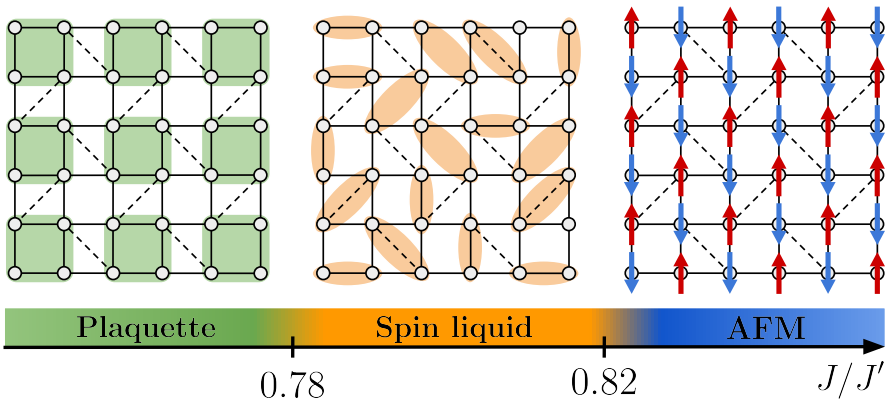


Figure 5.1: The ground-state phase diagram of the Shastry-Sutherland model as obtained in this work. The super-exchanges J and J' are denoted by solid and dashed lines, respectively.

approximation predicted an intermediate helical phase between the dimer and the Néel phases [170], while other works suggested a direct transition between these two phases [40, 171]. Later, an intermediate phase with plaquette order has been found by series expansion approaches [172] and confirmed within the generalization to $Sp(2N)$ symmetry and large-N expansion [173], by exact diagonalizations, and a combination of dimer- and quadrumer-boson methods [174]. Subsequent tensor-network approaches have corroborated the presence of the plaquette phase, for $0.675 \lesssim J/J' \lesssim 0.765$ [175]. This phase breaks the reflection symmetry across the lines containing the J' bonds (leading to a two-fold degenerate ground state) and is described by resonating singlets on half of the plaquettes with no J' bonds, see Fig. 5.1. The stabilization of plaquette order in $SrCu_2(BO_3)_2$ has been obtained when hydrostatic pressure is applied, even though there is evidence that the broken symmetry is related to the fourfold rotations around the center of plaquettes with no J' bonds [176, 177]. In addition, high-pressure thermodynamics provided evidence of a deconfined quantum critical point between the Néel and plaquette phases [178]. The latter aspect has been supported by a numerical analysis, also suggesting the emergence of the $O(4)$ symmetry at the critical point [179, 180]. However, recent density-matrix renormalization group (DMRG) and exact diagonalization calculations [181, 182] pushed forward the idea that a spin liquid intrudes between the antiferromagnetic and plaquette phases, around $0.79 \lesssim J/J' \lesssim 0.82$. The existence of an intruding spin-liquid phase has been also suggested by renormalization group calculations [183].

Numerical methods have proven crucial to obtain a description of the physical properties of the Shastry-Sutherland model or, in general, of other complicated physical systems. For example, the description of quantum states able to re-

produce the main features of quantum spin liquids is based on the concept of resonating-valence bond states [184, 185], leading to powerful physically inspired wave functions [41, 186, 187]. Although the construction of this kind of wave functions is generalizable to different models, it is not easy to define a systematic way to improve it; as a result, it is not always possible to achieve high accuracies for a generic model. On the other hand, DMRG and tensor-network approaches have also proved to be very competitive on two-dimensional systems [188, 189]. Still, despite a great computational effort, two-dimensional systems remain very challenging to deal with.

In this chapter, we aim to push the boundaries of Neural Quantum States by demonstrating that the ViT *Ansatz* introduced in Chapter 4, exclusively reliant on neural networks, enables us to achieve unprecedented accuracy in solving the challenging Shastry-Sutherland model. This model poses a particularly demanding problem in the realm of highly-frustrated magnetism, and our approach facilitates the extraction of its intricate physical properties. Carrying out simulations on clusters with periodic boundary conditions, we show that there exists a small, but finite, region in the phase diagram in which both the antiferromagnetic and plaquette order parameters vanish in the thermodynamic limit (see Fig. 5.1). As a result, this region is consistent with the existence of a spin-liquid state. Another original contribution of our studies is to define a suitable modification of the ViT architecture to treat excited states at finite momenta. This approach lends support to the existence of a *gapless* spin liquid.

5.2 Numerical results

We consider $L \times L$ clusters in which sites $\mathbf{r} = a/2(x, y)$ are labelled by x and y that take values from 0 to $L - 1$, and periodic-boundary conditions along the primitive vectors $\mathbf{a}_1 = (a, 0)$ and $\mathbf{a}_2 = (0, a)$, with $a = 2$. As a result, the total number of sites is L^2, while the number of unit cells is $L^2/4$. Most of the calculations are performed on clusters with L ranging from 6 to 14; in addition, calculations with $L = 16$ and 18 have been considered for $J/J' = 0.8$, located in the middle of the putative spin-liquid region. We mention that the ViT wave function breaks the spin $SU(2)$ symmetry and, therefore, it is not an eigenstate of the total spin S^2; still, it is possible to fix the z-component of the total spin, by performing the Monte Carlo sampling within a sector with a fixed value of $S^z = 1/2 \sum_i \sigma_i^z$. All calculations for assessing ground-state properties are performed taking $S^z = 0$ (see subsections 5.2.1 and 5.2.2). By contrast, triplet excitations are obtained by choosing $S^z = 1$ (see subsection 5.2.3). We emphasize that, the optimized states have extremely small deviations from the expected value of the total spin, e.g., $S(S+1) \approx 0.002$ (0.1) and 2.002 (2.07) for the singlet and triplet states for $L = 6$

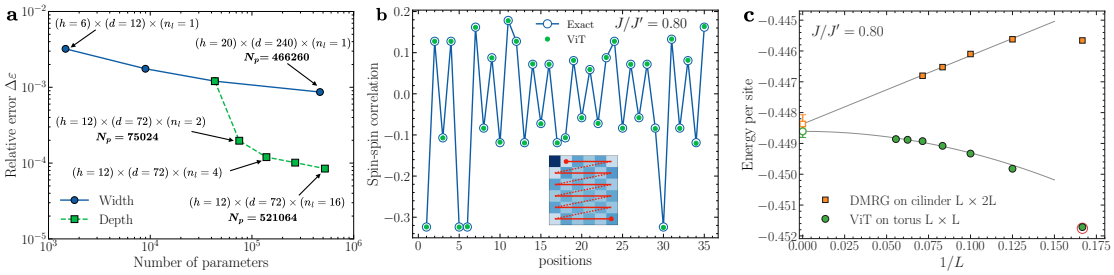


Figure 5.2: **Panel a:** Relative error $\Delta\varepsilon = |(E_{\text{exact}} - E_{\text{ViT}})/E_{\text{exact}}|$ of the ViT wave function on a 6×6 lattice at $J/J' = 0.8$. First, fixing only one layer and measuring the accuracy by increasing the width (blue dots). Then, for a fixed width, by increasing the number of layers (green squares). **Panel b:** The isotropic spin-spin correlations in real space as computed by the ViT wave function (full dots) on a 6×6 lattice at $J/J' = 0.8$. Values from exact diagonalization (empty dots) are also shown for comparison. Inset: The red line shows how the spin-spin correlations are ordered in the panel (b). **Panel c:** The comparison between the energies per site obtained by the ViT wave function (green circles) on $L \times L$ lattices with periodic-boundary conditions from $L = 6$ to $L = 18$ and the ones obtained by DMRG (orange squares) on $2L \times L$ cylinders with open-boundary conditions along the x direction from $L = 6$ to $L = 14$ [181]. The exact result on the 6×6 lattice is denoted with an empty red circle. The reported energy values were obtained by optimizing a ViT model with hyperparameters set to $h = 12$, $d = 72$, and $n_l = 8$, utilizing a sample size of $M = 6 \times 10^3$ during the optimization process.

$(L = 10)$ cluster, respectively.

5.2.1 Benchmarks

In order to validate our approach, we compare the results obtained by the ViT wave function with those obtained by exact diagonalizations on a small 6×6 cluster. Specifically, we focus on the challenging point $J/J' = 0.8$. We first examine the accuracy of the variational energies while varying the hyperparameters of the neural network. In Fig. 5.2a, we present the relative energy error as a function of the number of parameters, distributed in two different ways within the architecture. Initially, we maintain a single layer ($n_l = 1$) and increase the number of heads h and embedding dimension d. Subsequently, we fix a specific width ($h = 12$ and $d = 72$) and increment the number of layers from $n_l = 2$ to $n_l = 16$ (the energies are reported in Appendix 8.4.2).

This analysis highlights the importance of the model depth: for a fixed number of parameters, architectures that allocate parameters across multiple layers exhibit

superior accuracy. These results align with previous works [22, 23, 37, 133, 134], which underscore the necessity of deep neural networks for achieving high-precision results in two-dimensional frustrated systems. In addition, in Fig. 5.2b we show the isotropic spin-spin correlation functions $\langle \hat{\mathbf{S}}_\mathbf{0} \cdot \hat{\mathbf{S}}_\mathbf{r} \rangle$, illustrating that our variational wave function not only yields accurate energies, but also faithfully captures correlation functions at all distances. For cluster sizes exceeding $L = 6$, exact results become unattainable. Consequently, in Fig. 5.2c, we compare the variational energies of the ViT *Ansatz* on $L \times L$ clusters (with periodic-boundary conditions) to the ones obtained using the DMRG method on $L_x \times L_y$ cylinders with open and periodic boundaries in the x and y direction, respectively ($L_x = 2L_y$ and $L_y = L$ are considered) [181]. The energy per site is extrapolated in the thermodynamic limit, using system sizes ranging from $L = 8$ to $L = 18$ for the ViT wave function, and from $L = 8$ to $L = 14$ for the DMRG. To enhance the efficiency of the ViT for larger systems, specifically at $L = 16$ and $L = 18$, we employ a *local* attention mechanism (see Appendix 8.4.3 for further details). We mention that the energies obtained by the ViT wave function reveal a $1/L^2$ term as the leading correction, whereas the DMRG results exhibit an additional $1/L$ term. Most importantly, the energy extrapolated in the thermodynamic limit is compatible within the two approaches.

5.2.2 Phase diagram

Having proved the high accuracy of our *Ansatz*, we now focus on the region $0.7 \leq J/J' \leq 0.9$, which is expected to include both antiferromagnetic and plaquette phases, as well as the putative spin-liquid one. The presence of antiferromagnetic order is extracted from the thermodynamic limit of the staggered magnetization $m^2(L) = S(\pi, \pi)/L^2$ [181], where

$$S(\mathbf{k}) = \sum_{\mathbf{r}} e^{i\mathbf{k}\cdot\mathbf{r}} \langle \hat{\mathbf{S}}_\mathbf{0} \cdot \hat{\mathbf{S}}_\mathbf{r} \rangle \tag{5.2}$$

is the spin structure factor. Since the antiferromagnetic order pertains to the square lattice denoted by the sites $\mathbf{r}$, i.e., *without* considering the basis of the Shastry-Sutherland lattice, it is useful to define the momenta within this convention, i.e., $\mathbf{k} = 2\pi/L(n, m)$ with n and m taking values from 0 to $L - 1$. The existence of Néel order is signalled by a diverging peak at $\mathbf{k_m} = (\pi, \pi)$. In addition, the insurgence of the plaquette order is detected by a suitably defined order parameter

$$m_p(L) = |C(L/2, L/2) - C(L/2 - 1, L/2 - 1)| \ , \tag{5.3}$$

where the function $C(\mathbf{r})$ is defined as follows: starting from the operator $\hat{P}_\mathbf{r}$, which performs a cyclic permutation of the four spins of a plaquette with the top-right

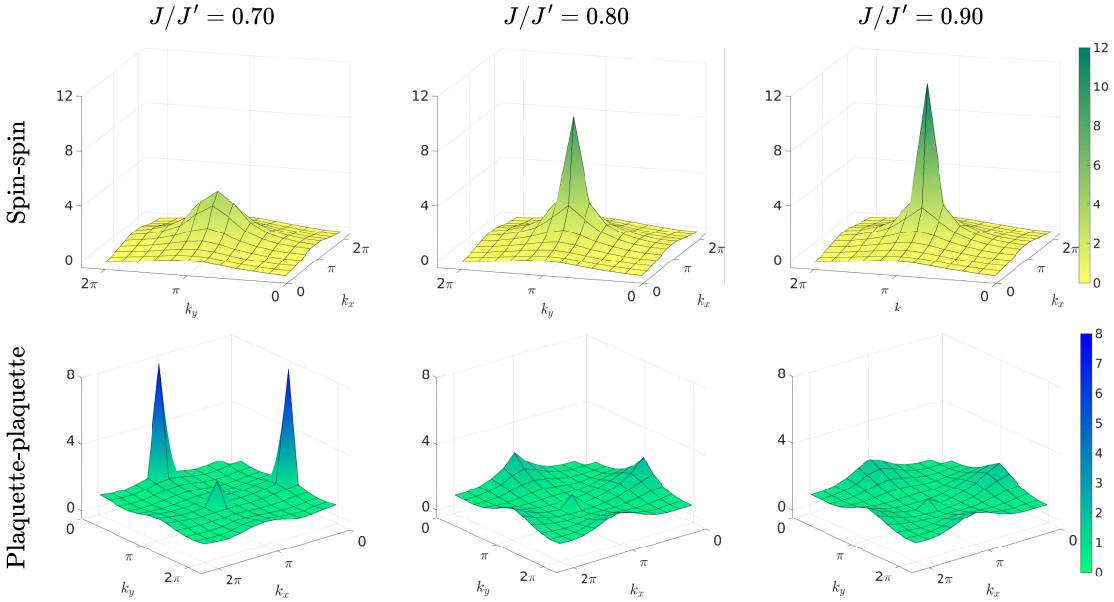


Figure 5.3: Fourier transform of the spin-spin (upper panels) and plaquette-plaquette (lower panels) correlations for $L = 12$ for different values of the frustration ratio J/J'. The calculations are performed with a Vision Transformer characterized by a number of heads equal to $h = 12$, an embedding dimension $d = 72$, and number of layers $n_l = 8$.

site at $\mathbf{r}$ [181], the following correlation functions are evaluated:

$$C(\mathbf{r}) = \frac{1}{4} \langle [\hat{P}_{\mathbf{r}} + \hat{P}_{\mathbf{r}}^{-1}][\hat{P}_{\mathbf{0}} + \hat{P}_{\mathbf{0}}^{-1}] \rangle \ . \tag{5.4}$$

Therefore, the plaquette order parameter $m_p(L)$ of Eq. (5.3) measures the difference, along the diagonal, of the plaquette correlation at the maximum distance and the second maximum distance; whenever the plaquette order is present, the correlation along the diagonal does not decay to zero, implying a non-vanishing value of $m_p(L)$ for large L. Similarly, the Fourier transform of the correlation functions in Eq. (5.4) (with the same conventions as for spins) denoted by $C(\mathbf{k})$ can be analyzed. The presence of the plaquette order can be identified by a diverging peak at $\mathbf{k_p} = (0, \pi)$ or $(\pi, 0)$. The results for $L = 12$ are shown in Fig. 5.3, for three values of the frustration ratio: for $J/J' = 0.7$ the ground state has strong peaks in $C(\mathbf{k})$ and a rather smooth spin structure factor $S(\mathbf{k})$, which is typical of a state with plaquette order; by contrast, for $J/J' = 0.9$ there are strong spin-spin correlations and weak plaquette-plaquette ones, which is characteristic of antiferromagnetic states.

In between, for $J/J' = 0.8$, the spin-spin correlations still have a peak, with

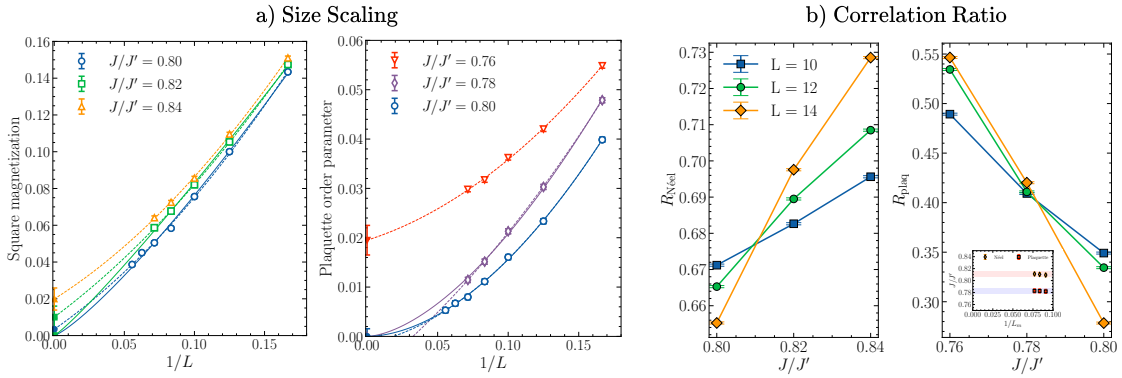


Figure 5.4: **Panel a:** Size scaling of the square magnetization $m^2(L)$ (left panel), and the plaquette order parameter $m_p(L)$ (right panel) as a function of $1/L$ from $L = 6$ up to $L = 18$. The values reported for each size L are obtained by extrapolating to an infinite number of layers (see Appendix 8.4.2 for the details), except for $L = 16$ and $L = 18$, where only simulations with $n_l = 8$ layers have been performed. The error bars of the extrapolated values in the thermodynamic limit are estimated via a resampling technique with gaussian noise. The fits associated to dashed curves are obtained using second-order polynomials in $1/L$ (see Eq. (5.5) of the main text), while solid curves are obtained using the critical form in Eq. (5.6) of the main text. **Panel b:** In the left (right) panel we show the correlation ratio $R_{\text{Néel}}$ (R_{plaq}) for the antiferromagnetic (plaquette) order in the interval $J/J' \in [0.80, 0.84]$ ($J/J' \in [0.76, 0.80]$). System sizes from $L = 10$ to $L = 14$ are used, and correlation ratio values are computed exclusively with architectures having $n_l = 8$ layers. Inset: Crossing points of the correlation ratio for Néel (orange diamond) and plaquette (red squares) order parameter as a function of the system size. The crossing points are obtained using $L_1 \times L_1$ and $L_2 \times L_2$ clusters with $(L_1, L_2) = (10, 12), (10, 14), (12, 14)$, with $L_m = (L_1 + L_2)/2$. Error bars on the correlation ratio are determined using resampling techniques under the assumption of Gaussian noise.

moderate plaquette correlations. In order to get information on the thermodynamic limit, a size scaling is necessary. In general, if magnetic order is stabilized, the square magnetization scales asymptotically as [190, 191]:

$$m^2(L) \approx m_0^2 + \frac{A_1}{L} + \frac{A_2}{L^2} \ , \tag{5.5}$$

where m_0 is the magnetization in the thermodynamic limit. In a disordered phase, the magnetization vanishes in the thermodynamic limit. The size corrections can be either exponential (for a gapped state) or power law (for a gapless one). In the vicinity of the Néel transition, the gap is relatively small and we use the "critical"

form [20]:

$$m^2(L) \approx L^{-(1+\eta)} \ . \tag{5.6}$$

Similar scaling behaviors are considered for $m_p(L)$ (within the plaquette phase, exponential corrections should be present, but no appreciable differences in the fits are observed with respect to the choice of a polynomial fit). In Fig. 5.4a, we perform a size-scaling extrapolation of both order parameters. For $J'/J = 0.84$ ($J'/J = 0.76$), the numerical values of the square magnetization (plaquette order parameter) fit well with a second-order polynomial in $1/L$ and suggest the existence of long-range order in the thermodynamic limit. By contrast, for $J'/J = 0.78, 0.8$ ($J'/J = 0.82, 0.84$), a more appropriate description of the scaling behavior of m^2 (m_p) is obtained by the critical relation of Eq. (5.6) (more details about the extrapolations are reported in the Appendix 8.4.2). This fact is compatible with the existence of a *gapless* spin liquid, which is also corroborated by the direct computation of the spin gap (see subsection 5.2.3). Interestingly, fitting the data of the square magnetization at $J'/J = 0.8$ with $m^2 \approx L^{-(1+\eta)}$, we get $\eta \approx 0.3$, in agreement with the DMRG calculations of Ref. [181]. We emphasize that, for the most challenging point $J/J' = 0.8$, lattices with $L = 16$ and 18 have been also considered, giving further support in favor of an intermediate spin liquid phase.

In summary, we find that the magnetization (plaquette order) vanishes for $J/J' \approx 0.82$ ($J/J' \approx 0.77$). These results suggest that a spin liquid exists between $(J/J')_{\text{plaq}} \approx 0.77$ and $(J/J')_{\text{Néel}} \approx 0.82$. To further support the present outcome, we measure the correlation ratio for the plaquette order as $R_{\text{plaq}} = 1 - C(\mathbf{k_p} + \delta\mathbf{k})/\mathbf{C}(\mathbf{k_p})$, and for the magnetic order as $R_{\text{Néel}} = 1 - S(\mathbf{k_m} + \delta\mathbf{k})/\mathbf{S}(\mathbf{k_m})$, where $||\delta\mathbf{k}|| = 2\pi/L$. When plaquette (magnetic) order is not present, $C(\mathbf{k})$ ($S(\mathbf{k})$) is a smooth function of $\mathbf{k}$, which implies that $R_{\text{plaq}} \to 0$ ($R_{\text{Néel}} \to 0$) in the thermodynamic limit; instead, when plaquette (magnetic) order settles down, $C(\mathbf{k})$ ($S(\mathbf{k})$) is finite for all the momenta except for $\mathbf{k_p}$ ($\mathbf{k_m}$), leading to $R_{\text{plaq}} \to 1$ ($R_{\text{Néel}} \to 1$). Then, the transition point may be accurately determined by locating the crossing point of the correlation ratio curves for different system sizes. The results for the plaquette (magnetic) order are shown in Fig. 5.4b, in the relevant interval $J/J' \in [0.76, 0.80]$ ($J/J' \in [0.80, 0.84]$), increasing the system size, i.e., for $L = 10$, 12, and 14. The various curves cross at $(J/J')_{\text{plaq}} \approx 0.78$ ($(J/J')_{\text{Néel}} \approx 0.81$), validating the phase boundary derived from the extrapolations of the order parameters.

5.2.3 Nature of the Spin Liquid

A crucial step toward understanding the nature of the spin liquid-phase is the characterization of its low-energy excitations. The first important question is to determine whether the energy spectrum is gapped or gapless. Here, we focus

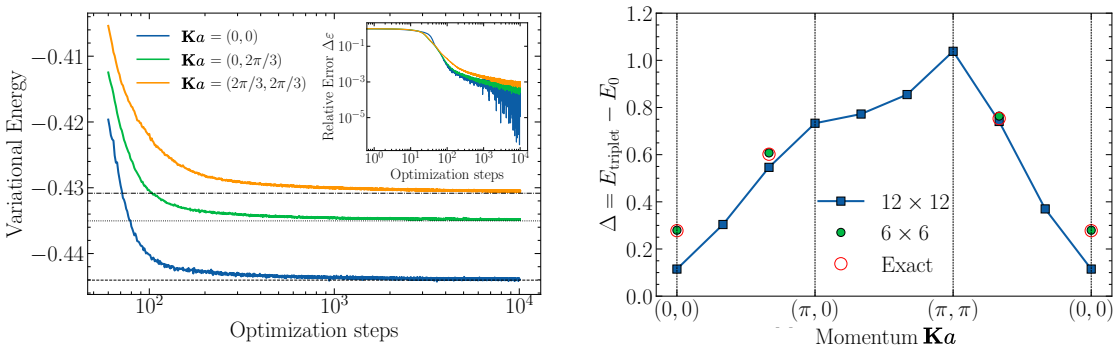


Figure 5.5: **Left panel:** Variational energy as a function of the optimization steps for the triplet momentum-resolved excitation spectra on a 6×6 lattice, with $\mathbf{K}a = (0,0)$, $(0, 2\pi/3)$, and $(2\pi/3, 2\pi/3)$; depicted in blue, green, and orange, respectively. The exact energies per site for the corresponding states are denoted by dashed ($E = -0.444040$), dotted ($E = -0.435033$), and dash-dotted ($E = -0.430831$) lines. The inset displays the relative error of the variational energy compared to exact diagonalization results. **Right panel:** The energy gap Δ between the ground state and the triplet excitations for the independent momenta of 6×6 lattice (green points), with the corresponding values obtained from exact diagonalization (red empty points). The energy gaps for a 12×12 lattice on a closed path in the Brillouin zone are denoted by blue squares. All the calculations are performed at a frustration ratio of $J/J' = 0.8$. The hyperparameters for these calculations are $h = 12$, $d = 72$, and $n_l = 8$. For the optimization protocol, we used a sample size of $M = 2^{14}$, a learning rate $\tau = 0.02$, and a diagonal shift regularization term of $\lambda = 10^{-3}$ (see Section 4.2).

on the frustration ratio $J/J' = 0.8$, a representative point within the spin-liquid phase, and assess triplet excitations for different momenta. Then, we focus on the lowest-energy state, which lies at the Γ point of the Brillouin zone, and perform the extrapolation to the thermodynamic limit. Our results provide strong evidence for a *gapless* spin liquid.

The ViT architecture outlined in Section 4.1.1 employs translationally invariant attention weights and the input patches are constructed from spins within the unit cell (see Appendix 8.4.1 for additional details). Consequently, the resulting wave function preserves translational symmetry with zero momentum. Nevertheless, by exploiting the translational equivariance property of the mapping between the input vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and output vectors $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, the architecture can be easily adapted to study an arbitrary sector with momentum $\mathbf{K}$, satisfying the following relation:

$$\mathrm{Log}[\Psi_\theta^K(T_{\mathbf{R}}\sigma)] = \mathrm{Log}[\Psi_\theta^K(\sigma)] + i\mathbf{K}\cdot\mathbf{R}\ , \tag{5.7}$$

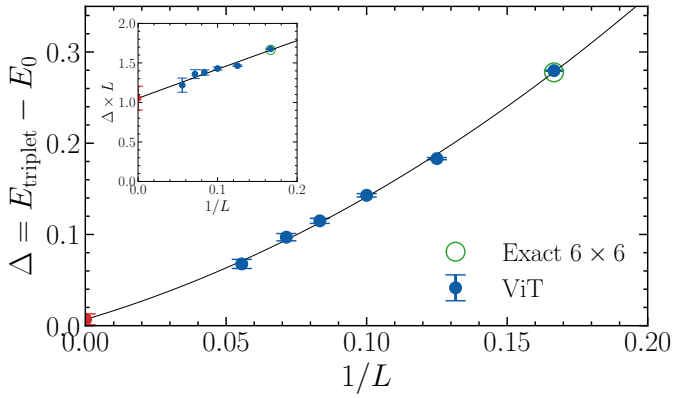


Figure 5.6: Energy gap Δ obtained by the ViT wave function between the ground state and the lowest-energy triplet state, as a function of inverse linear length $1/L$, from $L = 6$ to $L = 18$ at $J/J' = 0.8$. The exact gap for the 6×6 lattice is also reported for comparison (green empty circle). Inset: Rescaled gap $\Delta \times L$ as a function of $1/L$ for the same size considered in the main panel.

where $T_{\mathbf{R}}$ represents a translation by the Bravais lattice vector $\mathbf{R}$, whose components are integer multiples of the primitive vectors $\mathbf{a}_1$ and $\mathbf{a}_2$. The momenta $\mathbf{K}$ are quantized in units of $2\pi/L$, i.e, $K_x = 2\pi n/L$, with n ranging from 0 to $L/2-1$ and similarly for K_y. To define the ViT state with a specific momentum $\mathbf{K}$, we modified the amplitude as follows:

$$\text{Log}[\Psi_\theta^K(\sigma)] = \text{Log}[\Psi_\theta(\sigma)] + i\Theta_K(\sigma) \ , \tag{5.8}$$

where the function $\Theta_K(\sigma)$ adjusts the phase to match the target momentum sector, ensuring that Eq. (5.7) is satisfied. This phase shift can be computed from the output vectors $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ as follows:

$$\Theta_K(\sigma) = \text{Im}\left\{\text{Log}\left(\sum_{j=1}^{n} e^{i\mathbf{K}\cdot\mathbf{R}_j} y_{j,1}\right)\right\} \ . \tag{5.9}$$

where $\mathbf{R}_j$ indicates the (Bravais) vector that identifies the j-th patch of the Shastry-Sutherland model and $y_{j,1}$ is the first component of the vector $\boldsymbol{y}_j$, chosen by convention.

In practice, we fix one of the possible momenta allowed in the $L \times L$ lattice and perform the Monte Carlo sampling in the sector with $S^z = 1$. Even though the latter condition does not imply that the variational state is a triplet, we verified that the expectation value of the total spin S^2 is very close to 2. Therefore, this *Ansatz* gives an accurate approximation of a triplet state. The left panel of Fig. 5.5 shows the optimization curves of the variational energy for the three independent

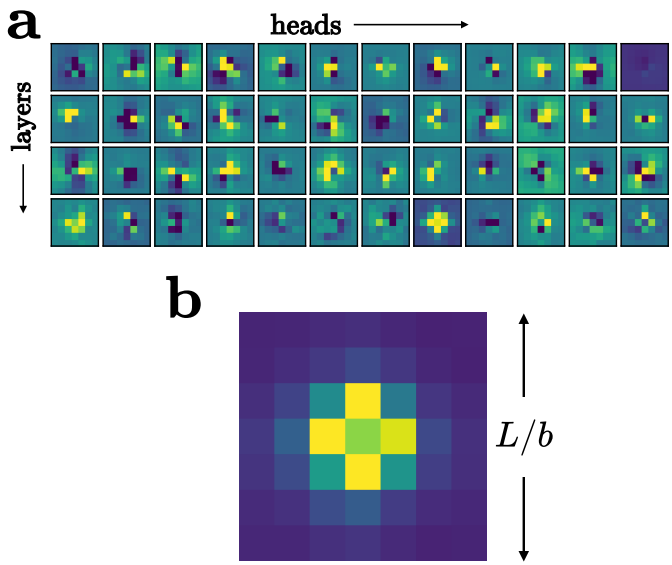


Figure 5.7: a) Attentions maps of a ViT with 4 layers and 12 heads per layer, optimized in the point $J/J' = 0.8$ for $L = 14$. b) Mean of the absolute values of all the attention maps. The attention maps have size L/b, where L and b are the linear dimensions of the lattice and of the patches, respectively.

momenta of the 6×6 cluster, namely $\mathbf{K}a = (0, 0)$, $(0, 2\pi/3)$, and $(2\pi/3, 2\pi/3)$ (where $a = 2$ is the length of the primitive vectors, to have $-\pi < K_x a \leq \pi$, and similarly for K_y). In all the cases, the ViT wave function achieves a relative error of approximately $\Delta\varepsilon \approx 10^{-3}$ within 10^4 optimization steps. The resulting variational triplet energy gaps have high accuracy when compared with the exact ones (refer to the right panel of Fig. 5.5). Specifically, we obtain that the lowest-energy excitation on the 6×6 lattice lies at the Γ point. In the right panel of Fig. 5.5, we also report the triplet gaps for a closed path in momentum space on a 12×12 cluster, confirming that the zero-momentum excitation remains the lowest-energy one.

Then, we proceed to analyze the size scaling of the lowest-energy triplet at $J/J' = 0.8$. The energy gap $\Delta = E_{\text{triplet}} - E_0$ is reported in Fig. 5.6 for different values of the cluster size, i.e., with L ranging from 6 to 18. The extrapolation performed using a quadratic fit of the form $\Delta = a + b/L + c/L^2$ yields, with a small fitting error, a vanishing gap in the thermodynamic limit, i.e., $\Delta = 0.00(7)$. In addition, in the inset of Fig. 5.6, we show that the rescaled gap $\Delta \times L$ approaches a finite value in the thermodynamic limit, thus confirming the gapless nature of the intermediate spin-liquid phase.

5.2.4 Attention maps

An approach to understand how the ViT wave function processes the input spin configurations is to examine the attention weights α_{i-j} of a trained Transformer for the different heads and layers at $J/J' = 0.8$ and $L = 14$, which we show in Fig. 5.7a. A key feature of the self-attention mechanism is to connect all the input patches even in a single layer. We highlight that the network makes use of this capability even in the first layer, since some heads attend to all the patches. This is not possible when working with architectures that use only local filters (e.g., convolutional ones). To capture the overall behavior of the attention weights, we compute the mean of the absolute values of the weights across all heads and layers. The results, displayed in Fig. 5.7b, reveal a systematic trend: the mean interaction between patches (representing groups of spins in this context) exhibits a regular decay as the relative distance between patches increases. Interestingly, this mean attention map encodes also the rotational symmetry of the model, which is not imposed in the weights, whereas individual attention maps do not exhibit this feature (see Fig. 5.7a). These observations are fully consistent with the findings of Ref. [139], in which analytical results establish a direct link between attention weights and correlations among spins.

5.3 Discussion

Our results demonstrate that NQSs represent an extremely useful tool to investigate the ground-state properties of frustrated quantum magnets. In this chapter, we focused the attention on the Shastry-Sutherland model, for which the existence of a spin-liquid phase between the plaquette and antiferromagnetic ones has been recently suggested [181, 182]. The difficulty of the problem resides in the smallness of this region, thus requiring extremely accurate calculations and large system sizes. The ViT wave function introduced in Chapter 4 (that combines a real-valued attention mechanism and a final complex-valued fully-connected layer) allows us to detect the existence of a finite region $0.78 \lesssim J/J' \lesssim 0.82$ in which both magnetic and plaquette orders vanish in the thermodynamic limit, thus supporting the presence of the intermediate spin-liquid phase [181]. Most importantly, by a direct evaluation of the energy gap between the ground state and the lowest-energy triplet excitation, we provided evidence for a *gapless* spin-liquid phase.

Chapter 6

The role of Queries and Keys in Neural Quantum States

In this chapter, we further benchmark the idea of factored attention, introduced in Chapter 1, in the framework of Neural Quantum States. In particular, we explore the suitability of the dot product attention mechanism, the cornerstone of modern Transformers, in the specific domain of the parametrization of variational wave functions to approximate ground states of quantum many-body spin Hamiltonians. The dot product attention mechanism was originally designed for natural language processing tasks to adeptly capture semantic relationships between word pairs in sentences by computing a similarity overlap between queries and keys. 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 a lattice. By comparing the performance of standard attention mechanisms with the simplified factored version that excludes queries and keys, relying solely on positions (see Chapter 1), 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.

6.1 The dot product attention mechanism

The success of Transformers lies in their remarkable flexibility: with minimal modifications, they excel in addressing diverse problem domains, often outperforming specialized approaches [7, 50, 51]. This is a consequence of their versatile founda-

tional components, namely the dot product self-attention mechanism, Multilayer Perceptron (MLP), Layer Normalization, and skip connections (see also Chapter 4). 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 [192, 193]. One wonders if the dot product attention mechanism provides an inductive bias which is the most appropriate in *any* data domain. For example, Ref. [70] in the context of protein contact prediction and Ref. [194] in computer vision tasks suggest that input-independent attention weights achieve competitive performance compared to the standard approach. In this chapter, we delve into this aspect by exploring the application of the Transformer architecture as a Neural Quantum State for approximating the ground state of quantum many-body spin Hamiltonians on lattice [17]. The Transformer architecture has been recently employed in this context, achieving highly accurate results across different systems [21, 37, 71, 113, 195–198]. While many of these works adopt the standard attention mechanism [113, 196, 197], in Section 4.3 with a simplified version, omitting queries and keys, we have reached *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 chapter, 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:

1. 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.

2. By analyzing the attention maps produced by architectures including queries and keys, we find that the optimization process makes them efficaciously input-independent.

3. 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.

6.2 Relative positional attention mechanisms

The success of the Transformer architecture is commonly attributed to the attention mechanism [1]. As already described in Chapter 1 and Chapter 3, the basic idea of the attention mechanism is to process an input sequence of n vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, with $\boldsymbol{x}_i \in \mathbb{R}^d$, producing a new sequence $(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$, with $\boldsymbol{A}_i \in \mathbb{R}^d$. The goal of this transformation is to construct context-aware output vectors by combining all input vectors [1]:

$$\boldsymbol{A}_i = \sum_{j=1}^{n} \alpha_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j) V \boldsymbol{x}_j \ . \tag{6.1}$$

The attention weights $\alpha_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 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 [201–203]. Here, we consider three different mechanisms, all based on relative positional encoding [201], as appropriate for the objective of this work.

1. *T5 attention*, introduced in Ref. [199], is one of the most popular attention mechanisms:

$$\alpha_{ij}^{T5}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\exp\left(\frac{\boldsymbol{x}_i^T Q^T K \boldsymbol{x}_j}{\sqrt{d}} + p_{i-j}\right)}{\sum_{k=1}^{n} \exp\left(\frac{\boldsymbol{x}_i^T Q^T K \boldsymbol{x}_k}{\sqrt{d}} + p_{i-k}\right)} \ . \tag{6.2}$$

2. *Decoupled attention*, introduced in Ref. [200]:

$$\alpha_{ij}^{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\exp\left(\frac{\boldsymbol{x}_i^T Q^T K \boldsymbol{x}_j}{\sqrt{d}}\right)}{\sum_{k=1}^{n} \exp\left(\frac{\boldsymbol{x}_i^T Q^T K \boldsymbol{x}_k}{\sqrt{d}}\right)} + p_{i-j} \ . \tag{6.3}$$

3. *Factored attention*, introduced in Refs. [70, 194] and in Chapter 1:

$$\alpha_{ij}^{F}(\boldsymbol{x}_i, \boldsymbol{x}_j) = p_{i-j} \ . \tag{6.4}$$

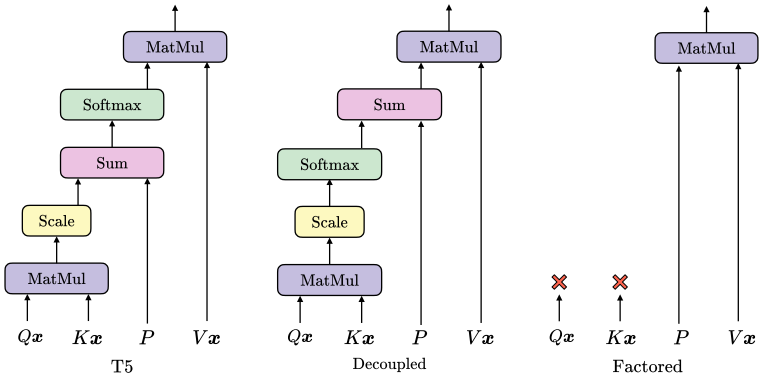


Figure 6.1: Schematic representation of the attention mechanisms employed in this work: T5 [199] (left panel), Decoupled [200] (central panel) and Factored [70, 91] (right panel) attention. In each of them, relative positional encoding is used. The matrices Q, K, V and P are referred to queries, keys, values and positional encoding matrix, respectively. Refer to Eqs. (6.2),(6.3) and (6.4) in the main text for the analytical expressions.

The vectors $Q\boldsymbol{x}_i$, $K\boldsymbol{x}_i$ and $V\boldsymbol{x}_i$ 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 Fig. 6.1, 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 Eq. (6.3), is the simplest extension of the Factored version in Eq. (6.4), 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 Eq. (6.2)] all the attention weights are constrained to be positive due to the global softmax activation.

6.3 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:

$$\hat{H} = J_1 \sum_{\langle i,j \rangle} \hat{\boldsymbol{S}}_i \cdot \hat{\boldsymbol{S}}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \hat{\boldsymbol{S}}_i \cdot \hat{\boldsymbol{S}}_j \ , \tag{6.5}$$

where $\hat{\boldsymbol{S}}_i = (S_i^x, S_i^y, S_i^z)$ 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 [190, 191]. 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 [164]. The determination of the precise nature of the ground state in the frustrated region remains challenging and subject to debate [20, 160, 204].

We employ a ViT wave function (see Section 4.1) to approximate, in the VMC framework, 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 6.2, 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 Eq. (6.2)], Decoupled [see Eq. (6.3)], and Factored [see Eq. (6.4)]. In Fig. 6.2, 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 [21, 37, 71], 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 Fig. 6.3, 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 $L = 6$ [205] and through variance extrapolation from Ref. [160] and Ref. [22], for $L = 8$ and $L = 10$, respectively. This plot demonstrates that the accuracy remains size-consistent across the tested

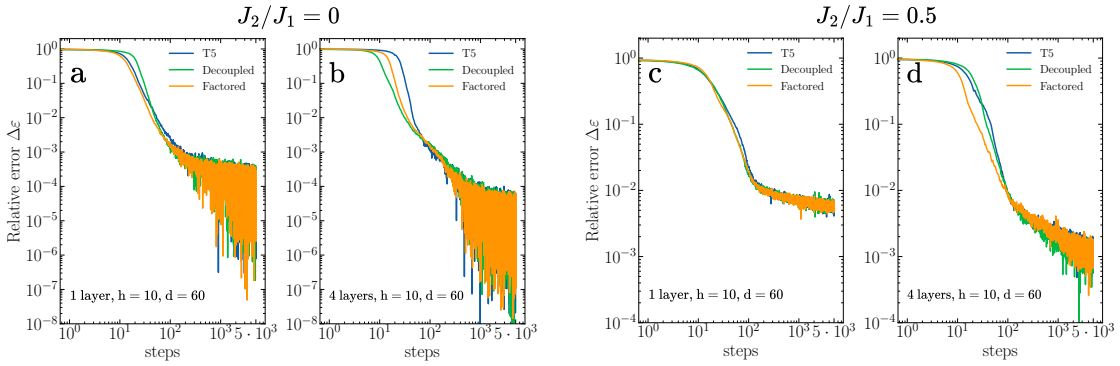


Figure 6.2: Relative error $\Delta\varepsilon = |(E_0 - E_{\text{ViT}})/E_0|$ during the optimization of the ViT wave function on the J_1-J_2 Heisenberg model at $J_2/J_1 = 0$ (left panel) and at $J_2/J_1 = 0.5$ (right) on a 6×6 lattice with periodic boundary conditions. The exact energies E_0 are computed with exact-diagonalization approaches. The architectures used for the simulations have $h = 10$ heads, embedding dimension $d = 60$, linear patch size $b = 2$, $n_l = 1$ layer in panels (a),(c), and $n_l = 4$ layers in panels (b),(d). All networks are trained with the same optimization protocol, using SR (see Section 4.2) for 5×10^3 optimization steps with $M = 6 \times 10^3$ samples for the stochastic estimates. A cosine decay learning rate scheduler is applied, starting with an initial value of $\tau = 0.03$. The optimization curves are consistent across multiple runs with different random initialization of the parameters.

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 6.1 we report the results on a 10×10 lattice at $J_2/J_1 = 0.5$, obtained using a four-layer architecture. The first column shows the final mean energy

	Energy	Parameters	Time
T5	-0.497025(6)	184,900	28h
Decoupled	-0.497108(6)	184,900	28h
Factored	-0.497184(6)	155,620	12.5h

Table 6.1: 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 10×10 lattice.

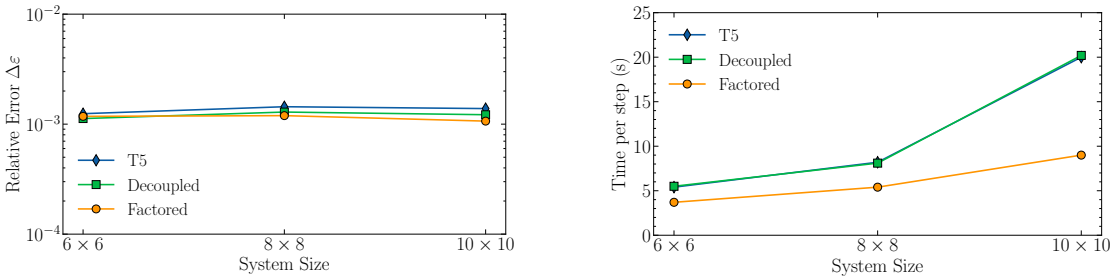
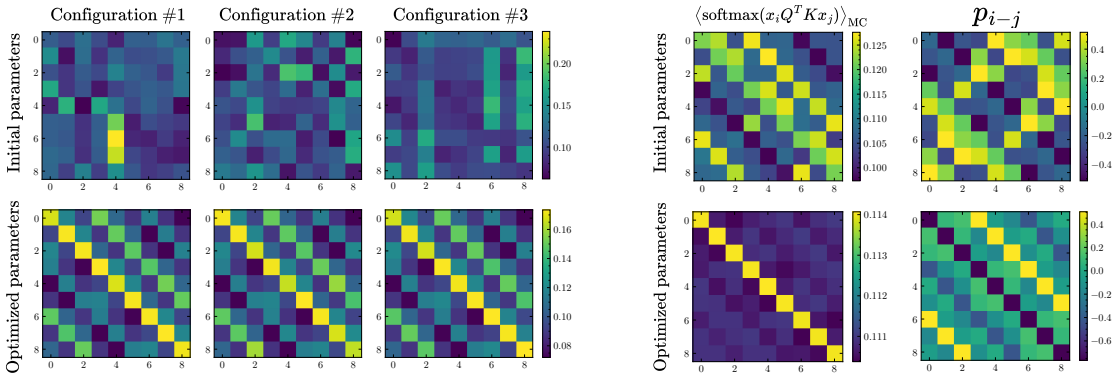


Figure 6.3: **Left panel:** Relative error $\Delta\varepsilon = |(E_0 - E_{\text{ViT}})/E_0|$ at $J_2/J_1 = 0.5$ as a function of the system size for ViT architectures with the three different attention mechanisms, namely Factored (orange circles), Decoupled (green squares) and T5 (blue diamonds). The reference ground state energies are taken from exact diagonalization for $L = 6$ (-0.503810) [205] and from variance extrapolation for $L = 8$ (-0.49906) [160] and $L = 10$ (-0.497715) [22]. **Right panel:** Time per optimization step (in seconds) measured on a single GPU A100 for the three attention mechanisms as a function of the system size. For all simulations a ViT architecture with hyperparameters $d = 10$, $h = 10$, $b = 2$ and $n_l = 4$ is considered. The model is optimized using the SR optimization method for 5×10^3 steps, employing $M = 6 \times 10^3$ training samples. A cosine decay learning rate scheduler is applied, starting with an initial value of $\tau = 0.03$.

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 the physical symmetries of the model through quantum number projection approaches [206, 207]; however, this goes beyond the scope of our work.

6.3.1 Analysis of the attention maps

The main result of the numerical simulations reported in Figs. 6.2, 6.3 and discussed in section 6.3 is that, using a ViT employing T5, Decoupled and Factored attention, the final accuracy is practically the same (see Table 6.1). 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 Fig. 6.4, we consider



a) T5 attention **b) Decoupled attention**

Figure 6.4: **Panel a:** Visualizations of the attention maps of a ViT with T5 attention mechanism [see Eq. (6.2)] for three different input spin configurations. When using initial random parameters there is a clear input dependence in the attention maps (top row). Instead, at the end of the optimization, the attention maps are practically input independent (bottom row). **Panel b:** Visualizations of the input-dependent term (left panels) and of the input-independent term (right panels) of a ViT with Decoupled attention mechanism [see Eq. (6.3)]. After the optimization (bottom row), the input-dependent term is approximately the identity matrix shifted element-wise by a constant, thus Factored attention is recovered [see Eq. (6.4)]. In the plots, the input-dependent term has been averaged over $M = 6 \times 10^3$ input configurations sampled from the optimized state. The presented results are obtained by optimizing a ViT architecture with a single layer $n_l = 1$, embedding dimension $d = 60$ and $h = 10$ different heads on a 6×6 lattice at $J_2/J_1 = 0.5$ (see panel (c) of Fig. 6.2). The linear patch size is taken to be $b = 2$, thus we have $n = 9$ patches and the resulting attention maps have shape 9×9. The plots are obtained by averaging the attention weights over all heads.

the case of T5 attention, plotting the attention weights defined in Eq. (6.2) 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 Fig. 6.4, we consider the case of Decoupled attention, plotting separately the input dependent and the positional contributions of the attention weights [see Eq. (6.3)]. 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

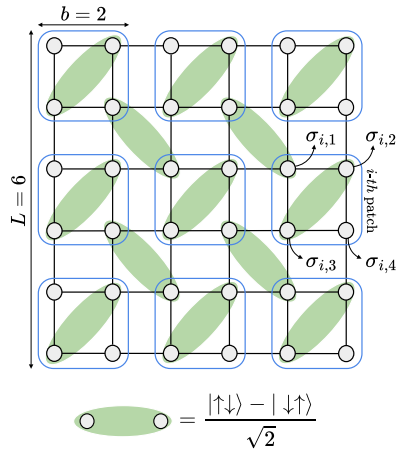


Figure 6.5: Graphical representation of the ground state of the Shastry-Sutherland model in the dimer phase [38] on a 6×6 lattice (periodic boundary connections not shown for clarity). The green shaded regions denote singlet states between two next-nearest neighbors spins. The blue squares $b \times b$ indicate the patches used to construct the input set of vectors for the Transformer.

words, Factored attention is spontaneously recovered from the Decoupled version (see section 6.2).

6.3.2 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 Eq. (6.2) and Eq. (6.3)] 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 [38], which captures the low-temperature properties of $SrCu_2(BO_3)_2$, a compound known for its intriguing physical properties [208]. 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 [38], refer to Fig. 6.5 for a graphical representation. Here, we want to show that a single-layer ViT with Factored attention [see Eq. (6.4)] 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 Fig. 6.5), 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 $\boldsymbol{x}_i = (\sigma_{i,1}, \sigma_{i,2}, \sigma_{i,3}, \sigma_{i,4})^T$, where $i = 1, \ldots, n$, with $n = L^2/b^2$. Then, we apply the Multi-Head attention mechanism [1] with $h = 2$ heads. Considering the value matrices:

$$V^{(1)} = \begin{pmatrix} 0 & 0 & 0 & V_{11}^{(1)} \\ 0 & V_{22}^{(1)} & V_{23}^{(1)} & 0 \end{pmatrix} \quad V^{(2)} = \begin{pmatrix} V_{14}^{(2)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \tag{6.6}$$

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

$$\boldsymbol{v}_i^{(1)} = \left(V_{11}^{(1)} \sigma_{i,4}, V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3} \right)^T \quad \boldsymbol{v}_i^{(2)} = \left(V_{14}^{(2)} \sigma_{i,1}, 0 \right)^T . \tag{6.7}$$

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:

$$S(i) = \begin{cases} (i+1)\%n & \text{if } \; i\%(L/b) = 0, \\ (i + L/b)\%n + 1 & \text{otherwise,} \end{cases} \tag{6.8}$$

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 Eq. (6.2)] or Decoupled [see Eq. (6.3)] attention mechanisms, we have to set $Q = K = 0$. The resulting attention vectors are:

$$\boldsymbol{A}_i^{(1)} = \left(V_{11}^{(1)} \sigma_{i,4}, V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3} \right)^T \quad \boldsymbol{A}_i^{(2)} = \left(V_{14}^{(2)} \sigma_{S(i),1}, 0 \right)^T . \tag{6.9}$$

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

$$W = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \tag{6.10}$$

we obtain:

$$\boldsymbol{A}_i = \left(V_{11}^{(1)} \sigma_{i,4} + V_{14}^{(2)} \sigma_{S(i),1}, V_{22}^{(1)} \sigma_{i,2} + V_{23}^{(1)} \sigma_{i,3}, 0, 0 \right)^T . \tag{6.11}$$

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 $F(\boldsymbol{A}_i+c)$, where c is a constant bias. The output of this operation is the sequence of vectors:

$$\boldsymbol{y}_i = \Big(F(V_{11}^{(1)}\sigma_{i,4} + V_{14}^{(2)}\sigma_{S(i),1} + c), F(V_{22}^{(1)}\sigma_{i,2} + V_{23}^{(1)}\sigma_{i,3} + c), 0, 0\Big)^T \ . \tag{6.12}$$

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

$$\boldsymbol{z} = \left(\sum_{i=1}^{n} F(V_{11}^{(1)}\sigma_{i,4} + V_{14}^{(2)}\sigma_{S(i),1} + c), \sum_{i=1}^{n} F(V_{22}^{(1)}\sigma_{i,2} + V_{23}^{(1)}\sigma_{i,3} + c), 0, 0\right)^T . \tag{6.13}$$

Replacing the fully-connected network that acts on $\boldsymbol{z}$ [21, 37, 141] with a simpler sum, we get the amplitude of the input spin configuration:

$$\mathrm{Log}[\Psi_\theta(\sigma)] = \sum_{i=1}^{n} \Big[F(V_{11}^{(1)}\sigma_{i,4} + V_{14}^{(2)}\sigma_{S(i),1} + c) + F(V_{22}^{(1)}\sigma_{i,2} + V_{23}^{(1)}\sigma_{i,3} + c)\Big] \ . \tag{6.14}$$

At the end, by choosing $F(\cdot) = \mathrm{logcos}(\cdot)$ 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 Fig. 6.5:

$$\Psi_0(\sigma) = \prod_{i=1}^{L^2/4} \cos\Big(\frac{\pi}{2} + \pi(\sigma_{i,4} + 3\sigma_{S(i),1})\Big) \cos\Big(\frac{\pi}{2} + \pi(\sigma_{i,2} + 3\sigma_{i,3})\Big) \ . \tag{6.15}$$

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* [209, 210]:

$$\lim_{|i-j|\to+\infty} \langle \hat{B}_i \hat{B}_j \rangle = \langle \hat{B}_i \rangle \langle \hat{B}_j \rangle \ , \tag{6.16}$$

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 Eq. (6.16) 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 [Eq. (6.2)] or Decoupled [see Eq. (6.3)] 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 Eq. (6.4)]. 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 Eq. (6.2) can provide a good inductive bias for solving the task.

6.4 Discussion

In this chapter, 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 introduced in Chapter 1. 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 [199]) should converge towards input-independent solutions.

Chapter 7

Foundation Neural-Networks Quantum States as a Unified Ansatz for Multiple Hamiltonians

In this chapter, inspired by the success of the Vision Transformer wave function developed in Chapter 4, we propose Foundation Neural-Network Quantum States (FNQS) as an integrated paradigm for studying quantum many-body systems. Foundation models are highly versatile neural-network architectures capable of processing different data types, such as text and images, and generalizing across various tasks like classification and generation. Similarly, FNQS leverage key principles of foundation models to define variational wave functions based on a single, versatile architecture that processes multimodal inputs, including spin configurations and Hamiltonian physical couplings. Unlike specialized architectures tailored for individual Hamiltonians, FNQS can generalize to physical Hamiltonians beyond those encountered during training, offering a unified framework adaptable to various quantum systems and tasks. FNQS enable the efficient estimation of quantities that are traditionally challenging or computationally intensive to calculate using conventional methods, particularly disorder-averaged observables. Furthermore, the fidelity susceptibility can be easily obtained to uncover quantum phase transitions without prior knowledge of order parameters. These pretrained models can be efficiently fine-tuned for specific quantum systems. The architectures trained in this chapter are publicly available at `https://huggingface.co/nqs-models`, along with examples for implementing these neural networks in NetKet.

7.1 Foundation models

The field of machine learning has undergone a fundamental transformation with the emergence of foundation models [211]. Built upon the Transformer architecture [1], these models have transcended their origins in language tasks [212, 213] to establish new paradigms across domains, from image generation [214] to protein structure prediction [215, 216]. Their efficacy emerges from a profound empirical observation: the scaling of models to hundreds of billions of parameters enables task-agnostic learning that achieves parity with specialized approaches while generating solutions for arbitrary problems defined at inference time [217]. These models exhibit remarkable generalization capabilities, enabling them to adapt to an extensive variety of tasks and domains without requiring task-specific fine-tuning. Another essential feature is their multimodality: they are trained on datasets comprising various formats, including text, images, videos, and audio, allowing them to process and generate outputs that combine these different forms. Foundation models have led to an unprecedented level of homogenization: almost all state-of-the-art natural language processing models are now adapted from a few foundation models. This homogenization produces extremely high leverage since enhancements to foundation models can directly and broadly improve performance across various applications.

In parallel, the study of quantum many-body systems has been significantly impacted by neural-network architectures employed as variational wave functions [17]. Neural-Network Quantum States (NQS) have emerged as a powerful framework for describing strongly-correlated models with unprecedented accuracy [20, 23, 25, 27, 218]. The modern formulations of the Stochastic Reconfiguration [108, 119, 138] method have enabled the stable optimization of variational states with millions of parameters [37, 219] (see Section 4.2), while the adaptation of the Transformer architecture for NQS parametrization [21, 71, 139, 141, 198, 220] presented in Chapter 4 has achieved state-of-the-art performance in challenging systems [21, 37] (see Section 4.3). Despite this progress, NQS are typically conceived in a system-specific fashion, and studying different Hamiltonians requires significant effort both in design and numerical optimization strategies.

To address these limitations, we present Foundation Neural-Network Quantum States (FNQS), a theoretical framework that synthesizes these advances by training neural-network-based variational wave functions capable of integrating as input not only the "standard" basis on which the wave function is represented, but also detailed information about the Hamiltonian (see Fig. 7.1). Our architecture is designed to achieve three key characteristics of foundation models in the quantum context: multimodality, through the ability to process multiple input types such as spin configurations and physical couplings; homogenization, by applying a single architecture across different Hamiltonians from simple to disordered systems; and

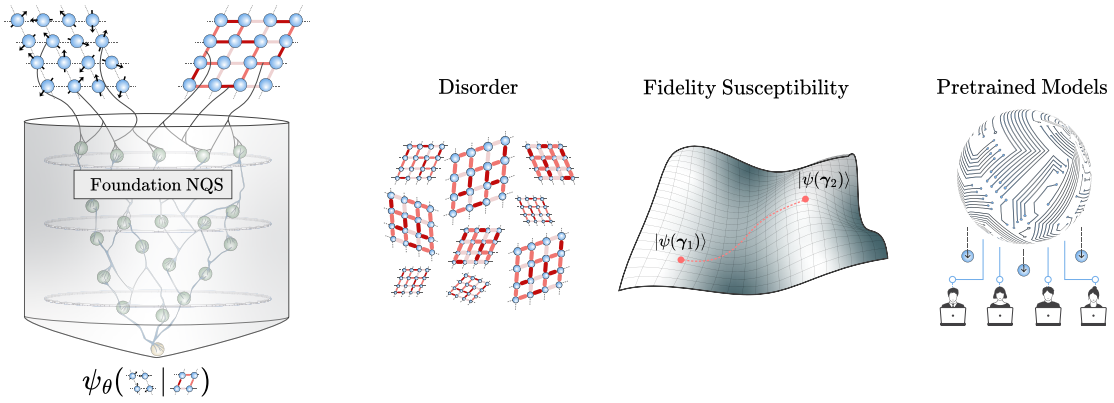


Figure 7.1: The first panel from the left shows a pictorial representation of Foundation Neural-Network Quantum States (FNQS), which, unlike traditional NQS, process multimodal inputs by incorporating both physical configurations and Hamiltonian couplings to define a variational wave function amplitude over their joint space. FNQS enable a range of applications, including the efficient simulation of disordered systems (see second panel) and the estimation of the quantum geometric tensor in coupling space, also known as the fidelity susceptibility, for the unsupervised detection of quantum phase transitions (see third panel). Moreover, FNQS combined with the public availability of the architectures allows users to leverage pretrained models to explore coupling regimes beyond those encountered during training (see forth panel).

generalization to physical Hamiltonians beyond the training dataset.

Previous efforts to construct foundation model-inspired wave functions have been reported in Refs. [221–225]. However, these approaches exhibit several limitations that are addressed in the present chapter. Specifically, some studies have been constrained to simple physical systems, achieving limited accuracy compared to specialized approaches [221], while others have employed ad hoc optimization strategies for chemical systems [222–224].

In contrast, we demonstrate applications that are unprecedented in both the diversity and complexity of physical models tackled by a single foundation model. We systematically explore systems of increasing complexity, including two-dimensional frustrated magnets with multiple couplings and disordered systems. This is enabled by the introduction of a suitably designed neural-network wave function based on the Transformer architecture developed in Chapter 4, combined with an optimization strategy that extends the Stochastic Reconfiguration method [108, 138] to simultaneously optimize across multiple systems. This generalized opti-

mization procedure is essential to achieve accurate results in the variational Monte Carlo framework.

Most notably, our framework enables simultaneous optimization of wave functions for multiple systems with computational complexity equivalent to single-system optimization, with no performance degradation as the number of systems increases. In addition, the framework enables efficient estimation of the fidelity susceptibility [43], providing rigorous, unsupervised detection of quantum phase transitions without prior knowledge of the order parameters [226, 227]. Refer to Fig. 7.1 for a pictorial representation of the different applications.

The chapter is organized as follows. In Section 7.2, we develop the theoretical framework for simultaneous training of variational wave functions across multiple quantum systems, adapting both Stochastic Reconfiguration for multi-system optimization and the Transformer architecture for multimodal quantum state parametrization. Section 7.3 presents systematic validation on the exactly solvable transverse field Ising model in one dimension, followed by an investigation of the J_1-J_2-J_3 Heisenberg model on a square lattice through fidelity susceptibility analysis. We conclude with an examination of disordered Hamiltonians, demonstrating the framework's capacity for efficient estimation of disorder-averaged quantities.

7.2 Theoretical framework for FNQS

The first step in developing foundation models to approximate ground states of quantum many-body Hamiltonians is to establish a theoretical framework that enables training a single NQS to approximate the ground states of multiple systems simultaneously. Consider a family of Hamiltonians, denoted by $\hat{H}_{\boldsymbol{\gamma}}$, where $\boldsymbol{\gamma}$ is a set of parameters that characterize each specific Hamiltonian, such as the physical couplings. Our goal is to find an approximation of the ground state of the ensemble of Hamiltonians $\hat{H}_{\boldsymbol{\gamma}}$ using a variational wave function $|\psi_\theta(\boldsymbol{\gamma})\rangle$ which explicitly depends on the physical couplings $\boldsymbol{\gamma}$ and on a shared set of variational parameters θ for all the Hamiltonians. To this end, we define the following loss function:

$$\mathcal{L}(\theta) = \int d\boldsymbol{\gamma}\, \mathcal{P}(\boldsymbol{\gamma}) \frac{\langle \psi_\theta(\boldsymbol{\gamma}) | \hat{H}_{\boldsymbol{\gamma}} | \psi_\theta(\boldsymbol{\gamma}) \rangle}{\langle \psi_\theta(\boldsymbol{\gamma}) | \psi_\theta(\boldsymbol{\gamma}) \rangle} \,, \tag{7.1}$$

where $\mathcal{P}(\boldsymbol{\gamma})$ is a normalized probability density over the couplings, i.e., $\int d\boldsymbol{\gamma} \mathcal{P}(\boldsymbol{\gamma}) = 1$. We denote expectation values with respect to the variational state $|\psi_\theta(\boldsymbol{\gamma})\rangle$ as $\langle \cdots \rangle_{\boldsymbol{\gamma}}$. This loss function represents an ensemble average of the energy expectation value $\langle \hat{H}_{\boldsymbol{\gamma}} \rangle_{\boldsymbol{\gamma}}$, weighted by the distribution $\mathcal{P}(\boldsymbol{\gamma})$. For each value of $\boldsymbol{\gamma}$, the variational energy $\langle \hat{H}_{\boldsymbol{\gamma}} \rangle_{\boldsymbol{\gamma}}$ is bounded from below by the exact ground state energy

$E_0(\boldsymbol{\gamma})$, such that $\langle \hat{H}_{\boldsymbol{\gamma}} \rangle_{\boldsymbol{\gamma}} \geq E_0(\boldsymbol{\gamma})$. Consequently, the loss function in Eq. (7.1) is bounded as $\mathcal{L}(\theta) \geq \mathcal{L}_0$, where $\mathcal{L}_0 = \int d\boldsymbol{\gamma} \mathcal{P}(\boldsymbol{\gamma}) E_0(\boldsymbol{\gamma})$ is the average ground state energy over the distribution $\mathcal{P}(\boldsymbol{\gamma})$.

The loss function in Eq. (7.1) can equivalently be written in a form amenable for Monte Carlo averages:

$$\mathcal{L}(\theta) = \int d\boldsymbol{\gamma}\, \mathcal{P}(\boldsymbol{\gamma}) \sum_{\boldsymbol{\sigma}} \frac{|\psi_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma})|^2}{\langle \psi_\theta(\boldsymbol{\gamma}) | \psi_\theta(\boldsymbol{\gamma}) \rangle} E_L(\boldsymbol{\sigma}, \boldsymbol{\gamma}) \ . \tag{7.2}$$

Here, we have introduced the local energy $E_L(\boldsymbol{\sigma}, \boldsymbol{\gamma}) = \langle \boldsymbol{\sigma} | \hat{H}_{\boldsymbol{\gamma}} | \psi_\theta(\boldsymbol{\gamma}) \rangle \,/\, \langle \boldsymbol{\sigma} | \psi_\theta(\boldsymbol{\gamma}) \rangle$ and the wave function $\langle \boldsymbol{\sigma} | \psi_\theta(\boldsymbol{\gamma}) \rangle = \psi_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma})$. The latter is parametrized by a neural network, and is the core variational object in our framework. Importantly, the explicit dependence of the many-body wave function amplitude $\psi_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma})$ on the Hamiltonian couplings $\boldsymbol{\gamma}$ is a major difference compared to traditional NQS and aligns with the principles of foundation models, where the capability to handle multiple data modalities, commonly referred to as *multimodality*, plays a central role (see Fig. 7.1).

The structure of the probability distribution $\mathcal{P}(\boldsymbol{\gamma})$ depends on the specific application. In disordered systems, a set of couplings $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_R\}$ can be directly sampled from $\mathcal{P}(\boldsymbol{\gamma})$, which may have continuous or discrete support. Conversely, in non-disordered systems, the probability distribution can be defined as $\mathcal{P}(\boldsymbol{\gamma}) = 1/\mathcal{R} \sum_{k=1}^{\mathcal{R}} \delta(\boldsymbol{\gamma} - \boldsymbol{\gamma}_k)$, where $\boldsymbol{\gamma}_k$ denotes the specific instances of the $\mathcal{R}$ Hamiltonians under study.

From a numerical perspective, an expectation value of a generic operator which is written in the form of Eq. (7.1), can be stochastically estimated using the Variational Monte Carlo framework [119], as discussed in Section 8.5.1 of the Appendix. In what follows, we denote by M the number of physical configurations used for the stochastic estimation of observables across $\mathcal{R}$ systems. Assuming that the samples are equally distributed across the systems, the number of samples per system is $M/\mathcal{R}$.

7.2.1 Stochastic Reconfiguration for multiple systems

A contribution of this chapter is the generalization of the Stochastic Reconfiguration (SR) method [108, 119, 138] to optimize variational wave functions that minimize loss of the form given in Eq. (7.1), thereby enabling the approximation of ground states across an ensemble of Hamiltonians. The key difference from the standard single-system problem is that, in this case, SR is derived as the minimizer of the ensemble-averaged fidelity between the exactly evolved state in imaginary time and the variationally evolved state (see Section 8.5.2 of Appendix for a detailed derivation).

The SR parameter updates $\delta\theta_\beta$ for $\beta = 1, \dots, P$, with P the total number of parameters, are obtained as

$$\boldsymbol{\delta\theta} = -\eta \, (\mathcal{S} + \lambda \mathbb{I}_P)^{-1} \boldsymbol{\mathcal{G}} \, , \tag{7.3}$$

where η is the learning rate and $\lambda > 0$ is a diagonal shift regularization to ensure the invertibility of the $\mathcal{S}$ matrix (see see Section 8.5.2 of Appendix). On the one hand, the gradient $\mathcal{G}_\alpha = \partial \mathcal{L}(\theta)/\partial\theta_\alpha$ is obtained as $\mathcal{G}_\alpha = \int d\boldsymbol{\gamma} \mathcal{P}(\boldsymbol{\gamma}) G_\alpha(\boldsymbol{\gamma})$ with

$$G_\alpha(\boldsymbol{\gamma}) = 2 \, \mathrm{Re} \left\{ \langle \hat{H}_{\boldsymbol{\gamma}} \hat{O}_{\boldsymbol{\gamma},\beta} \rangle_{\boldsymbol{\gamma}} - \langle \hat{H}_{\boldsymbol{\gamma}} \rangle_{\boldsymbol{\gamma}} \langle \hat{O}_{\boldsymbol{\gamma},\beta} \rangle_{\boldsymbol{\gamma}} \right\} \, . \tag{7.4}$$

Here, $\hat{O}_{\boldsymbol{\gamma},\alpha}$ is a diagonal operator in the computational basis of the system characterized by couplings $\boldsymbol{\gamma}$, whose matrix elements are defined as $\langle \boldsymbol{\sigma} | \hat{O}_{\boldsymbol{\gamma},\alpha} | \boldsymbol{\sigma}' \rangle = \delta_{\sigma,\sigma'} \partial \mathrm{Log}[\psi_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma})]/\partial\theta_\alpha$. On the other hand, the matrix $\mathcal{S}$ in the extended space is $\mathcal{S}_{\alpha\beta} = \int d\boldsymbol{\gamma} \mathcal{P}(\boldsymbol{\gamma}) S_{\alpha\beta}(\boldsymbol{\gamma})$ with

$$S_{\alpha\beta}(\boldsymbol{\gamma}) = \mathrm{Re} \left\{ \langle \hat{O}^\dagger_{\boldsymbol{\gamma},\alpha} \hat{O}_{\boldsymbol{\gamma},\beta} \rangle_{\boldsymbol{\gamma}} - \langle \hat{O}^\dagger_{\boldsymbol{\gamma},\alpha} \rangle_{\boldsymbol{\gamma}} \langle \hat{O}_{\boldsymbol{\gamma},\beta} \rangle_{\boldsymbol{\gamma}} \right\} \, . \tag{7.5}$$

Importantly, the matrix $\mathcal{S}$ is constructed as an ensemble average of the matrices $S_{\alpha\beta}(\boldsymbol{\gamma})$ of the individual systems, weighted by the probability distribution $\mathcal{P}(\boldsymbol{\gamma})$. In the absence of this weighting, $\mathcal{S}$ would reduce to an unweighted integral, leading to large statistical fluctuations and potentially diverging variances in its elements.

For simplicity, we have omitted the dependency of the log-derivative operator $\hat{O}_{\boldsymbol{\gamma},\alpha}$ and the $\mathcal{S}$ matrix on the variational parameters θ.

7.2.2 Foundation Neural-Network architecture

To parametrize the FNQS, we adapt the Vision Transformer (ViT) Ansatz introduced in Chapter 4 to process multimodal inputs, defined by the physical configurations $\boldsymbol{\sigma}$ and the Hamiltonian couplings $\boldsymbol{\gamma}$.

The traditional ViT architecture processes the physical configuration $\boldsymbol{\sigma}$ in three main steps (see Chapter 4 for a detailed description):

1. *Embedding.* The input configuration $\boldsymbol{\sigma}$ is split into n patches, where the specific shape of the patches depends on the structure of the lattice and its dimensionality, see for example [21, 71, 139]. Then, the patches are embedded in $\mathbb{R}^d$ through a linear transformation of trainable parameters, defining a sequence of input vectors $(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)$.

2. *Transformer Encoder.* The resulting input sequence is processed by a Transformer Encoder, which produces another sequence of vectors $(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n)$, with $\boldsymbol{y}_i \in \mathbb{R}^d$ for all i.

3. *Output layer.* These vectors are summed to produce the hidden representation $\boldsymbol{z} = \sum_{i=1}^{n} \boldsymbol{y}_i$, which is finally mapped through a fully-connected layer to a single complex number representing the amplitude corresponding to the input configuration. Only the parameters of this last layer are taken to be complex-valued.

The generalization of the architecture to include as inputs the couplings $\boldsymbol{\gamma}$ is performed by modifying *only* the *Embedding* step described above. In particular, we adopt two different strategies, which cover the systems studied in this chapter, depending on whether the parameter vector $\boldsymbol{\gamma}$ consists of $O(1)$ or $O(N)$ real numbers, with N indicating the total number of physical degrees of freedom of the model. We stress that the property of having a single, versatile architecture that can be adapted to study physical systems with distinct characteristics, such as a different number of couplings, is a key property of foundation models, also called *homogenization* (see Section 7.1). In the first scenario where the auxiliary parameters are $O(1)$, we concatenate the values of the couplings to each patch of the physical configuration before the linear embedding. Then the usual linear embedding procedure in $\mathbb{R}^d$ is performed. Instead, in the second scenario with $O(N)$ external parameters, we split the vector of the couplings into patches using the same criterion used for the physical configuration. We then use two different embedding matrices to embed the resulting patches of the configuration and of the couplings, generating two sequences of vectors: $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ with $\boldsymbol{x}_i \in \mathbb{R}^{d/2}$ for the physical degrees of freedom and $(\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n)$ with $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{d/2}$ for the couplings. The final input to the Transformer is constructed by concatenating the embedding vectors, forming the sequence $(\mathrm{Concat}(\boldsymbol{x}_1, \tilde{\boldsymbol{x}}_1), \ldots, \mathrm{Concat}(\boldsymbol{x}_n, \tilde{\boldsymbol{x}}_n))$, with $\mathrm{Concat}(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i) \in \mathbb{R}^d$. Notice that after the first layers, the representations of the configurations and of the couplings are mixed by the attention mechanism. The Embedding step can be generalized to any general parameterized Hamiltonian represented as a graph [228].

Regarding the lattice symmetries encoded in the architecture, for non-disordered Hamiltonians we employ a translationally invariant attention mechanism that ensures a variational state invariant under translations among patches [21, 139]. In contrast, for disordered models, we do not impose constraints on the attention mechanism.

7.2.3 Generalized Fidelity Susceptibility

A rigorous approach for the unsupervised detection of quantum phase transitions involves measuring the *fidelity susceptibility* [43]. Consider a general system described by the Hamiltonian $\hat{H}_{\boldsymbol{\gamma}}$ characterized by N_c couplings $\boldsymbol{\gamma} = (\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(N_c)})$.

We define the *generalized fidelity susceptibility* as:

$$\chi_{ij}(\boldsymbol{\gamma}) = -\left.\frac{\partial^2 \ln F(\boldsymbol{\gamma},\boldsymbol{\varepsilon})}{\partial\varepsilon_i\partial\varepsilon_j}\right|_{\boldsymbol{\varepsilon}=0} , \tag{7.6}$$

where $F(\boldsymbol{\gamma},\boldsymbol{\varepsilon})$ is the fidelity:

$$F^2(\boldsymbol{\gamma},\boldsymbol{\varepsilon}) = \frac{|\langle\psi_\theta(\boldsymbol{\gamma})|\psi_\theta(\boldsymbol{\gamma}+\boldsymbol{\varepsilon})\rangle|^2}{\langle\psi_\theta(\boldsymbol{\gamma})|\psi_\theta(\boldsymbol{\gamma})\rangle\,\langle\psi_\theta(\boldsymbol{\gamma}+\boldsymbol{\varepsilon})|\psi_\theta(\boldsymbol{\gamma}+\boldsymbol{\varepsilon})\rangle} . \tag{7.7}$$

The fidelity measures the overlap between two quantum states on the manifold of the couplings $\boldsymbol{\gamma}$ and it shows a dip in correspondence with a quantum phase transition [43, 226, 227]. In the case of a single coupling ($N_c = 1$), the tensor $\chi_{ij}(\boldsymbol{\gamma})$ simplifies to a scalar function, which peaks at the phase transition and diverges in the thermodynamic limit. However, even in this simpler case, computing the fidelity [see Eq. (7.7)] is challenging, as it becomes an exponentially small quantity with increasing system size. Typically, the fidelity susceptibility is determined using exact diagonalization techniques on small clusters or tensor network-based methods for one-dimensional systems [229]. These methods involve calculating the ground state for each coupling value and applying finite-difference techniques to estimate the second derivative [see Eq. (7.6)]. Efficient algorithms based on Quantum Monte Carlo methods have been proposed to address this challenge, but they are limited to systems with positive-definite ground states [43]. In this chapter, we propose an alternative approach that overcomes these limitations. Specifically, we start by rewriting the generalized fidelity susceptibility in Eq. (7.6) as the quantum geometric tensor with respect to the Hamiltonian couplings $\boldsymbol{\gamma}$ [226, 227], namely:

$$\chi_{ij}(\boldsymbol{\gamma}) = \mathrm{Re}\left\{\langle\hat{\mathcal{O}}^\dagger_{\boldsymbol{\gamma},i}\hat{\mathcal{O}}_{\boldsymbol{\gamma},j}\rangle_{\boldsymbol{\gamma}} - \langle\hat{\mathcal{O}}^\dagger_{\boldsymbol{\gamma},i}\rangle_{\boldsymbol{\gamma}}\langle\hat{\mathcal{O}}_{\boldsymbol{\gamma},j}\rangle_{\boldsymbol{\gamma}}\right\} , \tag{7.8}$$

which is a $N_c \times N_c$ symmetric positive-definite matrix. The operators $\hat{\mathcal{O}}_{\boldsymbol{\gamma},i}$ are diagonal in the computational basis whose matrix elements are defined as $\langle\boldsymbol{\sigma}|\hat{\mathcal{O}}_{\boldsymbol{\gamma},i}|\boldsymbol{\sigma}'\rangle = \delta_{\sigma,\sigma'}\partial\mathrm{Log}[\psi_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma})]/\partial\gamma^{(i)}$, where $\gamma^{(i)}$ is the i-th component of the coupling vector $\boldsymbol{\gamma}$ [refer to Section 8.5.5 of the Appendix for a derivation of the equivalence between Eq. (7.8) and Eq. (7.6)]. By exploiting the multimodal nature of the FNQS wave function, it is possible to compute the derivatives of the amplitudes with respect to the Hamiltonian couplings, a highly non-trivial quantity that is inaccessible for standard variational states optimized on a single value of the couplings. As a result, for FNQS, the quantum geometric tensor in Eq. (7.8) can be directly computed using automatic differentiation techniques, bypassing the need to explicitly calculate the fidelity [see Eq. (7.7)].

We emphasize that identifying quantum phase transitions without prior knowledge of order parameters is a challenging task, and existing state-of-the-art methods have notable limitations that hinder their applicability in complicated scenarios. For instance, supervised approaches [230] require prior knowledge of the different phases, while unsupervised techniques are generally restricted to models with a single physical coupling [231] or rely on quantum tomography, which is typically computationally demanding [232, 233]. All these limitations are overcome by our approach, which extends the computation of fidelity susceptibility [43] to general physical models with multiple couplings.

7.3 Numerical results

The framework introduced in Section 7.2 is general and applicable to quantum systems of different nature. However, in the following, we specialize to spin-1/2 systems.

7.3.1 Transverse field Ising chain

In the first place, we test the framework on the one-dimensional Ising model in a transverse field, an established benchmark problem of the field. The system is described by the following Hamiltonian (with periodic boundary conditions):

$$\hat{H} = -J \sum_{i=1}^{N} \hat{S}_i^z \hat{S}_{i+1}^z - h \sum_{i=1}^{N} \hat{S}_i^x \ , \tag{7.9}$$

where $\hat{S}_i^x$ and $\hat{S}_i^z$ are spin-1/2 operators on site i. The ground-state wave function, for $J, h \geq 0$, is positive definite in the computational basis, with a known exact solution. In this case, the Hamiltonian depends on a single coupling, specifically the external magnetic field h.

In the thermodynamic limit, the ground state exhibits a second-order phase transition at $h/J = 1$, from a ferromagnetic ($h/J < 1$) to a paramagnetic ($h/J > 1$) phase. In finite systems with N sites, the estimation of the critical point can be obtained from the long-range behavior of the spin-spin correlations, that is, $m^2(\boldsymbol{\gamma}) = 1/N \sum_{i=1}^{N} \langle \hat{S}_{\boldsymbol{\gamma},i}^z \hat{S}_{\boldsymbol{\gamma},i+N/2}^z \rangle_{\boldsymbol{\gamma}}$. The quantum phase transition at $h/J = 1$ is in the universality class of the classical two-dimensional Ising model [234].

Here, we first demonstrate the ability to train a FNQS across multiple Hamiltonians, and even across quantum phase transitions. To achieve this, we use the variational Ansatz described in Section 7.2.2, training it on a chain of $N = 100$ sites across five different values of the external field ($\mathcal{R} = 5$), including values representative of both the disordered ($h/J = 1.2, 1.1$) and the magnetically ordered

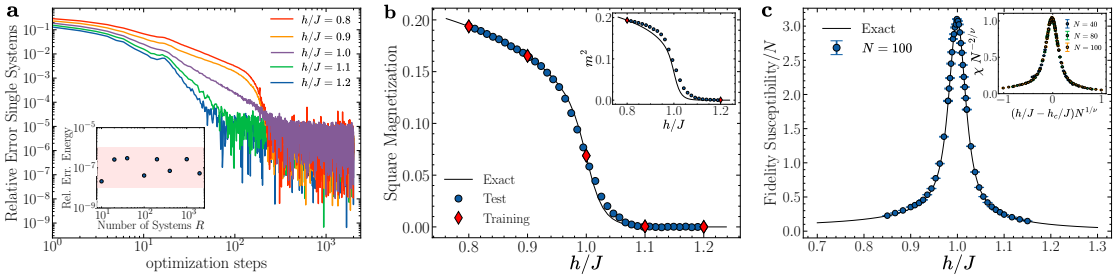


Figure 7.2: All the panels refer to the Ising model on a chain [see Eq. (7.9)]. **Panel a.** Simultaneous ground state energy optimization of $\mathcal{R} = 5$ systems on a chain of $N = 100$ sites, with external fields $h/J = 0.8, 0.9, 1.0, 1.1$ and 1.2. The relative error with respect to the exact ground state energy of each system is shown as a function of the optimization steps. The inset displays the relative error of the total energy as a function of the number of systems $\mathcal{R}$, defined by equispaced values of h/J in the interval $h/J \in [0.8, 1.2]$, with a fixed batch size of $M = 10000$. **Panel b.** Square magnetization evaluated with a FNQS trained at $h/J = 0.8, 0.9, 1.0, 1.1$ and 1.2 (red diamonds) and tested on previously *unseen* values of the external field (blue circles). The inset shows the square magnetization predictions of an architecture trained exclusively on $h/J = 0.8$ and 1.2, evaluated at intermediate external field values. **Panel c.** Fidelity susceptibility per site [see Eq. (7.8)] as a function of the external field for a FNQS trained on $\mathcal{R} = 6000$ equispaced values of h/J in the interval $h/J \in [0.85, 1.15]$ for a cluster of $N = 100$ sites. The inset shows the data collapse of the same quantity for $N = 40$, 80, and 100.

phase ($h/J = 0.9, 0.8$), as well as the transition point ($h/J = 1.0$). As shown in Fig. 7.2a, this single neural network describes all five ground states with high accuracy. The learning speed is only moderately different in the different states. In particular, the state with a value of h/J close to the transition point is the one that converges last. For the same architecture, we systematically vary the value of $\mathcal{R} \in [5, 2000]$, choosing the transverse field equispaced within the interval $h/J \in [0.8, 1.2]$. We keep the total batch size fixed to $M = 10000$, assigning an equal number of samples $M/\mathcal{R}$ across the $\mathcal{R}$ different systems. In the inset of panel (a), we show the relative error of the total energy accuracy as a function of $\mathcal{R}$. Remarkably, despite the number of systems increasing, the network's performance remains constant, with no observable degradation in accuracy. Crucially, this robustness is achieved at a *computational cost independent of the total number of systems*, as it depends solely on the neural network architecture and the fixed total batch size M. This result is a first illustration of the accuracy, scalability,

and computational efficiency of our approach.

Then, we investigate the generalization properties of the FNQS. In panel (b) of Fig. 7.2, we use the architecture trained with $\mathcal{R} = 5$ and evaluate its performance on external field values not included in the training set. In particular, we compute the square magnetization for other intermediate values of h/J, showing robust generalization capabilities of the network across the entire phase diagram. The inset explores a more restricted scenario in which training is performed using only two points: one in the disordered phase ($h/J = 1.2$) and another in the ordered phase ($h/J = 0.8$). This analysis shows that, even with minimal training data, the network avoids overfitting the ground state at these two points and learns a sufficiently smooth description of the magnetization curve.

Finally, in panel (c) of Fig. 7.2, we use a FNQS trained on $\mathcal{R} = 6000$ different points equispaced in the interval $h/J \in [0.85, 1.15]$ to calculate the fidelity susceptibility (see Section 7.2.3), comparing the FNQS results to the exact solution that is available in this case [229, 235]. In the inset of the same panel, we present a data collapse analysis of the fidelity susceptibility. Specifically, we show the scaled fidelity susceptibility $\chi N^{-2/\nu}$ versus $(h/J - h_c/J)N^{1/\nu}$ according to the scaling laws of Refs. [43, 226, 236, 237]. The data collapses well under $h_c/J = 1.00(1)$ and the critical exponent $\nu = 1.00(2)$ corresponding to the classical two-dimensional Ising universality class [238].

This first benchmark example highlights the ability of the FNQS to interpolate meaningfully between different phases, even when trained on a limited set of Hamiltonians. We attribute this capability to the properties of the ViT architecture employed. In particular, the multi-head attention mechanism could play a crucial role. For example, each attention head can, in principle, specialize in capturing features associated with distinct phases of the system. Moreover, the all-to-all connectivity intrinsic to the attention mechanism allows the network to flexibly describe long-range correlations, which are essential for accurately describing critical phenomena.

7.3.2 J_1-J_2-J_3 Heisenberg model

We now proceed to analyzing the J_1-J_2-J_3 Heisenberg model on a two-dimensional $L \times L$ square lattice with periodic boundary conditions:

$$\hat{H} = J_1 \sum_{\langle \boldsymbol{r},\boldsymbol{r'} \rangle} \hat{\boldsymbol{S}}_{\boldsymbol{r}} \cdot \hat{\boldsymbol{S}}_{\boldsymbol{r'}} + J_2 \sum_{\langle\langle \boldsymbol{r},\boldsymbol{r'} \rangle\rangle} \hat{\boldsymbol{S}}_{\boldsymbol{r}} \cdot \hat{\boldsymbol{S}}_{\boldsymbol{r'}} + J_3 \sum_{\langle\langle\langle \boldsymbol{r},\boldsymbol{r'} \rangle\rangle\rangle} \hat{\boldsymbol{S}}_{\boldsymbol{r}} \cdot \hat{\boldsymbol{S}}_{\boldsymbol{r'}} \ , \tag{7.10}$$

where $\hat{\boldsymbol{S}}_{\boldsymbol{r}} = (\hat{S}^x_{\boldsymbol{r}}, \hat{S}^y_{\boldsymbol{r}}, \hat{S}^z_{\boldsymbol{r}})$ represents the spin-1/2 operator localized at site $\boldsymbol{r}$; in addition, J_1, J_2, and J_3 are first-nearest-, second-nearest-, and third-nearest-neighbor antiferromagnetic couplings, respectively. The ground-state properties of this frustrated model have been extensively studied using various numerical and analytical

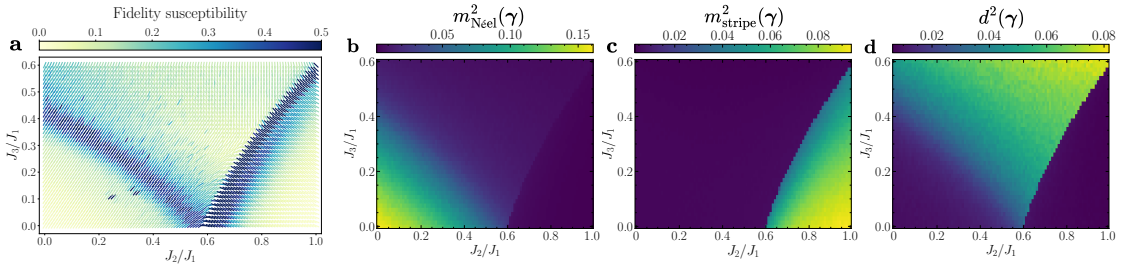


Figure 7.3: **Panel a.** Fidelity susceptibility of the J_1-J_2-J_3 Heisenberg model on a 10×10 square lattice [see Eq. (7.10)]. For each point of the phase diagram of the system, we visualize the direction of the leading eigenvector of the quantum geometric tensor $\chi_{ij}(\boldsymbol{\gamma})$ [see Eq. (7.8)]. The colour associated to each line is related to corresponding eigenvalue clipped in the interval $[0.0, 0.5]$ for visualization purposes. **Panel b.** The order parameter $m^2_{\text{Néel}}(\boldsymbol{\gamma})$ characterizing the Néel antiferromagnetic order. **Panel c.** The order parameter $m^2_{\text{stripe}}(\boldsymbol{\gamma})$ identifying the antiferromagnetic phase with stripe order. **Panel d.** The order parameter $d^2(\boldsymbol{\gamma})$ probing the valence bond phase. In all panels, the order parameters are computed over a dense grid of $\mathcal{R} = 4000$ uniformly distributed points in the parameter space defined by $J_2/J_1 \in [0, 1.0]$ and $J_3/J_1 \in [0, 0.6]$.

approaches. However, a complete characterization of its phase diagram remains challenging [239–246]. It is well established that antiferromagnetic order dominates in extended regions for $J_1 \gg J_2, J_3$ [with pitch vector $\boldsymbol{k} = (\pi, \pi)$] and for $J_2 \gg J_1, J_3$ [with pitch vectors $\boldsymbol{k} = (\pi, 0)$ or $\boldsymbol{k} = (0, \pi)$]. In contrast, in the intermediate region, frustration suppresses magnetic order, leading to valence-bond solid and, as recently suggested, spin-liquid states [245, 246]. The study of this model using FNQS aims to demonstrate that a single architecture can learn to effectively combine input spin configurations and Hamiltonian couplings, constructing a compact representation that captures and differentiates between distinct phases.

First, we aim for an initial characterization of the phase diagram in a fully unsupervised manner, aiming to distinguish regions with valence-bond ground states from those with magnetic order using the generalized fidelity susceptibility introduced in Section 7.2.3. To this end, we train a FNQS on a 10×10 lattice over a broad region of parameter space, setting a dense grid of $\mathcal{R} = 4000$ evenly spaced points in the plane defined by $J_2/J_1 \in [0, 1.0]$ and $J_3/J_1 \in [0, 0.6]$. Having two couplings J_2/J_1 and J_3/J_1, the quantum geometric tensor in the couplings space is a 2×2 matrix [see Eq. (7.8)]. For each point $\boldsymbol{\gamma} = (J_2/J_1, J_3/J_1)$ we diagonalize $\chi_{ij}(\boldsymbol{\gamma})$ and in Fig. 7.3a we visualize the direction of the eigenvector corresponding

to the maximum eigenvalue using lines, whose colors are associated to the leading eigenvalues and indicate the intensity of maximum variation of the variational wave function. We note that the lines of maximal variation partition the plane into three distinct regions, in agreement with the three different phases identified by the order parameters (see below). Remarkably, within this approach we are able to identify the existence of two phase transitions without any prior knowledge of the physical properties of the system. Furthermore, by analyzing the behavior of the eigenvectors, we can infer the nature of these phase transitions. For example, on the left branch of maximum variation, the eigenvectors exhibit no significant change in direction before and after the transition, which is indicative of a continuous phase transition. In contrast, the right branch shows a pronounced change in the eigenvector directions across the transition, suggesting a first-order phase transition. To the best of our knowledge, this is the first calculation of fidelity susceptibility for a system with more than one coupling. Indeed, without our approach, it would be highly computationally expensive to optimize thousands of systems with different coupling values, using finite difference methods to estimate the geometric tensor in the couplings space [see Eq. (7.8)].

To further analyze the physical property of the model, we compute the order parameters in each region of the phase diagram by examining spin-spin and dimer-dimer correlations. Specifically, for fixed values of the Hamiltonian couplings $\boldsymbol{\gamma} = (J_2/J_1,\ J_3/J_1)$, the antiferromagnetic order is detected by analyzing the spin structure factor

$$C(\boldsymbol{k};\boldsymbol{\gamma}) = \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \langle \hat{\boldsymbol{S}}_{\boldsymbol{\gamma},\boldsymbol{0}} \cdot \hat{\boldsymbol{S}}_{\boldsymbol{\gamma},\boldsymbol{r}} \rangle_{\boldsymbol{\gamma}} \ , \tag{7.11}$$

where $\boldsymbol{r}$ runs over all the lattice sites of the square lattice. On the one side, the antiferromagnetic Néel order is detected by measuring $m^2_{\text{Néel}}(\boldsymbol{\gamma}) = C(\pi,\pi;\boldsymbol{\gamma})/N$ [190, 191] with $N = L^2$. On the other side, the stripe antiferromagnetic order is identified by $m^2_{\text{stripe}}(\boldsymbol{\gamma}) = [C(0,\pi;\boldsymbol{\gamma}) + C(\pi,0;\boldsymbol{\gamma})]/(2N)$. In Section 8.5.4 of the Appendix, we measure these two order parameters by restricting the optimized FNQS to the axis $J_3 = 0$, allowing comparison with other techniques. Furthermore, the valence-bond solid order is detected by the dimer-dimer correlations:

$$\begin{aligned} D_{\alpha}(\boldsymbol{r};\boldsymbol{\gamma}) = 9\Big[& \langle \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{0}} \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{\alpha}} \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{r}} \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{r}+\boldsymbol{\alpha}} \rangle_{\boldsymbol{\gamma}} \\ & - \langle \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{0}} \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{\alpha}} \rangle_{\boldsymbol{\gamma}} \langle \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{r}} \hat{S}^z_{\boldsymbol{\gamma},\boldsymbol{r}+\boldsymbol{\alpha}} \rangle_{\boldsymbol{\gamma}} \Big] \ , \end{aligned} \tag{7.12}$$

where $\boldsymbol{\alpha} = \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$. Notice that the previous definition involves only the z component of the spin operators, which is sufficient to detect the dimer order [71, 130]; however, since we consider only one component, we include a factor of 9 in Eq. (7.12) to account for this [247]. Then, the corresponding structure factor is expressed

as $\mathcal{D}_\alpha(\boldsymbol{k};\boldsymbol{\gamma}) = \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} D_\alpha(\boldsymbol{r};\boldsymbol{\gamma})$. The order parameter to detect the valence-bond order is defined as $d^2(\boldsymbol{\gamma}) = \left[\mathcal{D}_x(\pi,0;\boldsymbol{\gamma}) + \mathcal{D}_y(0,\pi;\boldsymbol{\gamma})\right]/N$.

In panels (b), (c), and (d) of Fig. 7.3, we present the order parameters $m^2_{\text{Néel}}(\boldsymbol{\gamma})$, $m^2_{\text{stripe}}(\boldsymbol{\gamma})$, and $d^2(\boldsymbol{\gamma})$, which respectively characterize the antiferromagnetic Néel, antiferromagnetic stripe, and valence bond solid phases, as functions of the couplings $J_2/J_1 \in [0, 1.0]$ and $J_3/J_1 \in [0, 0.6]$. Comparing the different panels in Fig. 7.3, we observe a strong correspondence between the phase transition boundaries predicted by fidelity susceptibility and those identified through order parameters. This agreement validates our approach to the unsupervised detection of quantum phase transitions, even in systems with multiple couplings.

7.3.3 Random transverse field Ising model

A natural extension of this method involves exploring Hamiltonians with quenched disorder, by optimizing a single FNQS across distinct disorder realizations. Disordered systems are a very vast and ramified topic of research and are at the basis of a theory of complexity [248]. When quantum effects are also included, disordered systems become even more compelling, with recent works highlighting the extension of Anderson localization to a complete ergodicity breaking in interacting quantum systems [249]. These systems are notoriously resilient to numerical approaches [250] and optimizing a single FNQS across many realizations of disorder makes the averaging of the physical quantities, a necessary step for treating disordered systems, much more efficient [251].

A compelling candidate for study is the random transverse field Ising chain, defined by the following Hamiltonian (assuming periodic boundary conditions):

$$\hat{H} = -J \sum_{i=1}^{N} \hat{S}_i^z \hat{S}_{i+1}^z - \sum_{i=1}^{N} h_i \hat{S}_i^x \ , \tag{7.13}$$

where h_i is the on-site transverse magnetic field at the i-th site. In the disordered case, h_i varies randomly along the chain, drawn independently and identically from the uniform distribution on the interval $[0, h_0]$. When setting $J = 1/e$, the model exhibits a quantum phase transition between ordered (ferromagnetic) and disordered (paramagnetic) phases for $h_0 = 1$ [252–255]. Although this disordered model cannot be solved analytically due to the lack of translational symmetry, the eigenstates can be found efficiently for each realization of disorder by exploiting the mapping to free fermions [254]. Therefore, relatively large clusters may be considered, just requiring diagonalizations of $N \times N$ matrices [254]. This model is deceptively simple, since for a large region going from the critical point inside the disordered phase, it is affected by Griffiths-McCoy singularities [252, 253].

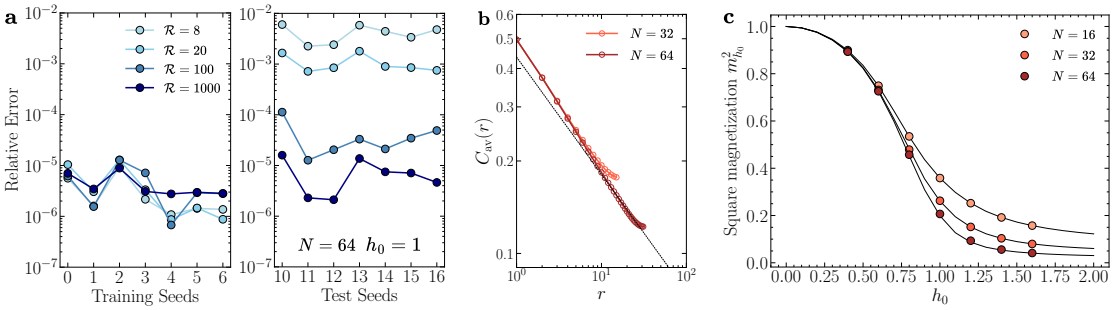


Figure 7.4: All the panels refer to the random transverse field Ising model on a chain [see Eq. (7.13)]. **Panel a.** Relative error of the variational energy on a cluster of $N = 64$ sites, fixing $h_0 = 1.0$ on different train (left) and test (right) disorder realizations increasing the number of systems $\mathcal{R}$. The integer numbers (seeds) shown on the x-axis label different disorder realizations. Specifically, integers from 0 to 6 correspond to a set of seven different training disorder realizations, while those from 10 to 16 correspond to a set of test realizations. **Panel b.** Spin-spin correlation function averaged over $\mathcal{R} = 1000$ disorder realizations at $h_0 = 1$. The dashed line represent the theoretical power law behaviour with exponent $\eta \approx 0.382$. **Panel c.** Square magnetization $m^2_{h_0}$ as a function of h_0. At fixed h_0 order parameter is obtained by averaging over $\mathcal{R} = 1000$ different disorder realizations. The numerically exact results are report as comparison with solid lines.

From a numerical perspective, unlike in previous cases, the coupling distribution $\mathcal{P}(\boldsymbol{\gamma})$ is a uniform distribution for the N transverse fields h_i in Eq. (7.13). Consequently, for each realization of disorder, the number of couplings is equal to the number of sites of the lattice (see Section 7.2.2 for more details about the strategies to treat this situation). This scenario provides an opportunity to assess the generalization capabilities of the neural network, particularly in its ability to accurately predict properties for new disorder realizations beyond those considered during the training.

In Fig. 7.4a, we optimize a single FNQS on a cluster of $N = 64$ sites. Training is carried out on $\mathcal{R}$ distinct disorder realizations, sampled by fixing $h_0 = 1$. The left (right) panel presents the relative error of the variational energy for seven different training (test) seeds as a function of the number of training realizations, namely $\mathcal{R} = 8, 20, 100, 1000$, while keeping in all cases the total batch size of spin configurations constant at $M = 10000$. The analysis reveals that increasing $\mathcal{R}$ does not compromise the accuracy on the training seeds. In fact, even with an increase in training points to $\mathcal{R} = 1000$, we achieve highly accurate energy predictions while keeping the number of configurations per system relatively low, specifically

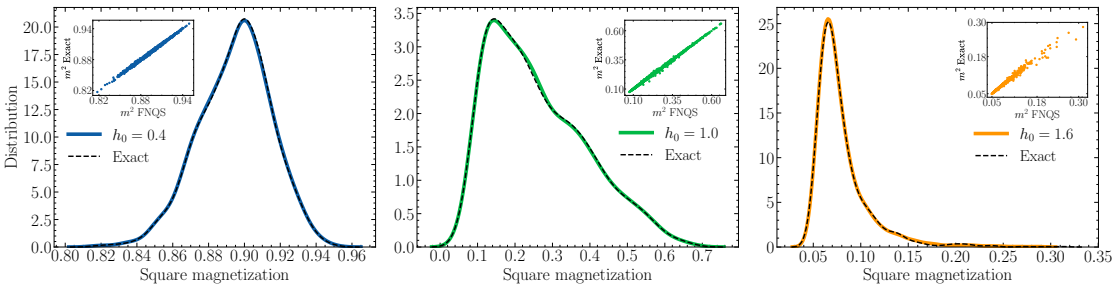


Figure 7.5: The distribution of the squared magnetization $m^2_{h_0}$ is analyzed for an FNQS trained on the random transverse field Ising model [see Eq. (7.13)] with chain length $N = 32$. The FNQS is trained on $\mathcal{R} = 1000$ independent disorder realizations and tested on a separate set of 1000 unseen realizations. The reported distributions correspond to the latter, with results presented for three distinct disorder strengths: $h_0 = 0.4$, $h_0 = 1.0$, and $h_0 = 1.6$. For comparison, numerically exact results are included as black dashed lines. The insets of each panel illustrate the correlation between the exact squared magnetizations and the variational values predicted by the FNQS for unseen disorder realizations.

$M/\mathcal{R} = 10$. More importantly, the generalization error on the test seeds (disorder realizations not encountered during training) systematically decreases when increasing $\mathcal{R}$. Notably, for $\mathcal{R} = 1000$, the relative errors of the training and test accuracies show the same order of magnitude, indicating that the FNQS has successfully learned how to combine the disorder couplings with the spin configurations to generate accurate amplitudes in the space of both physical configurations and couplings. We emphasize that the relative error for each disorder realization achieved by the FNQS is comparable to that obtained by training the same architecture on a single disorder realization (not reported here). This highlights the remarkable efficiency of the proposed method.

To assess the ability of FNQS to accurately predict disorder-averaged observables beyond energy, in Fig. 7.4b we show the average spin-spin correlation function at criticality:

$$C_{\text{av}}(r) = \frac{1}{N}\sum_{i=1}^{N}\int d\boldsymbol{\gamma}\mathcal{P}(\boldsymbol{\gamma})\,\langle \hat{S}^z_{\boldsymbol{\gamma},i}\hat{S}^z_{\boldsymbol{\gamma},i+r}\rangle_{\boldsymbol{\gamma}} \ . \tag{7.14}$$

The average correlation function $C_{\text{av}}(r)$ is stochastically estimated by sampling $\mathcal{R} = 1000$ disorder realizations at $h_0 = 1$. Refer to Section 8.5.1 of the Appendix for further details. We find good agreement with the theoretical critical scaling, characterized by the critical exponent $\eta = (3-\sqrt{5})/2 \approx 0.382$, which is depicted as a dashed line in Fig. 7.4b. In Fig. 7.4c we measure the order parameter of the

system as a function of h_0. In particular, for a fixed value of h_0, ranging from $h_0 = 0.4$ to $h_0 = 1.6$, we train a single FNQS over $\mathcal{R} = 1000$ distinct disorder realizations sampled for each h_0. After training, we estimate the square magnetization, defined as $m^2_{h_0} = 1/N \sum_{r=1}^{N} C_{\rm av}(r)$. The variational results are in excellent agreement with numerically exact calculations across different system sizes, namely $N = 16, 32, 64$. Remarkably, achieving similar results with standard methods would require the optimization of 1000 independent simulations for each value of h_0, highlighting the efficiency and scalability of our approach. To provide a more stringent test of the accuracy of the predicted observables, in Fig. 7.5 we analyze the distribution of the square magnetization $m^2_{h_0}$ over a set of 1000 *test* disorder realizations *not* encountered during training. The comparison with exact results demonstrates excellent agreement for the different values of $h_0 = 0.4, 1.0$ and 1.6, capturing not only the regions of high probability density but also the tails of the distributions with remarkable accuracy. In the inset of each panel of Fig. 7.5, we present correlation plots comparing the exact square magnetization with the FNQS predictions for disorder realizations *not* encountered during training. These plots further highlight the excellent agreement between the predictions and exact results, even for the most extreme and improbable values of the square magnetization.

7.3.4 Out of distribution generalization

In this section, we investigate whether a FNQS trained on a restricted coupling domain can generalize to couplings outside this domain, namely for $\boldsymbol{\gamma} \notin \mathrm{Dom}\{\mathcal{P}(\boldsymbol{\gamma})\}$. In general, we do not expect this type of out-of-distribution generalization to succeed, as in any other machine learning approach. Nonetheless, we report here an example where this type of unconventional generalization is effective with some limitations.

Specifically, we consider the following Hamiltonian defined by generalizing the J_1-J_2 Heisenberg model on a square lattice (with periodic boundary conditions) [20, 41]:

$$\begin{aligned}\hat{H} = J_1 \sum_i \left(\hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_{i+\hat{y}} + \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_{i+\hat{x}} \right) \\ + \sum_i \left(J_{2R}\ \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_{i+\hat{x}+\hat{y}} + J_{2L}\ \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_{i+\hat{x}-\hat{y}} \right)\end{aligned} \tag{7.15}$$

which depends on two distinct couplings, J_{2L}/J_1 and J_{2R}/J_1. When $J_{2L} = J_{2R} = 0$, the model reduces to the unfrustrated Heisenberg model on a square lattice, whose ground state exhibits antiferromagnetic order [190]. Increasing J_{2L} introduces frustration exclusively along the left diagonals of the square lattice, while increasing J_{2R} does so along the right diagonals (see Fig. 7.6a). In the limiting cases where

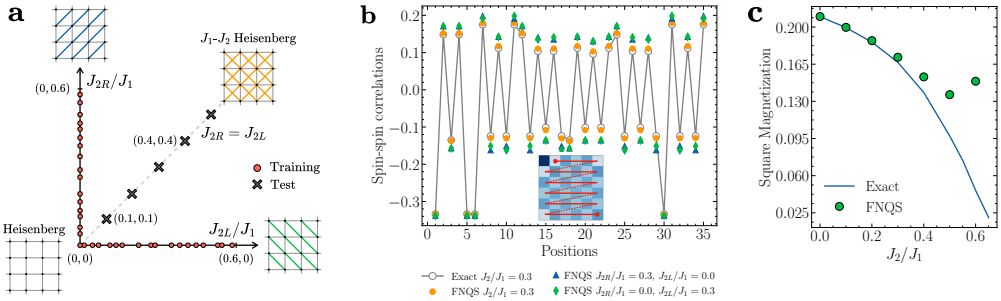


Figure 7.6: **Panel a:** Generalized J_1–J_2 Heisenberg model on a square lattice, with two couplings, J_{2L}/J_1 and J_{2R}/J_1. When $J_{2L} = J_{2R}$, the model reduces to the standard J_1–J_2 Heisenberg model. The FNQS is trained on configurations with frustration on only one diagonal and tested on configurations where both diagonals are frustrated. **Panel b:** Spin–spin correlation function on a 6×6 lattice of a FNQS trained on $\mathcal{R} = 1000$ different realizations of frustration affecting only one of the two diagonals of a square lattice (see Fig. 7.6 for a schematic representation). The model is tested on two in-distribution points $(J_{2L}/J_1, J_{2R}/J_1) = (0.0, 0.3)$ and $(0.3, 0.0)$, and on the out-of-distribution point $(J_{2L}/J_1, J_{2R}/J_1) = (0.3, 0.3)$, where both diagonals are frustrated. For reference, the exact results of the J_1–J_2 Heisenberg model at $J_2 = 0.3$ are also shown. Inset: The red line shows how the spin-spin correlations are ordered in the panel (b). **Panel c:** Out-of-distribution generalization of the square magnetization associated with Néel antiferromagnetic order for the FNQS (green circles) as a function of the frustration ratio J_2/J_1 on a 6×6 cluster. Exact results (blue line) are shown for reference.

either $J_{2L} \neq 0$ and $J_{2R} = 0$, or vice versa, the model in Eq. (7.15) corresponds to the Heisenberg model on the anisotropic triangular lattice [256–258].

To probe the generalization capability of the FNQS model, we design the following experiment (illustrated in Fig. 7.6a): the architecture is trained solely on Hamiltonians where frustration is present on only one diagonal at a time and then evaluated on Hamiltonians in which both diagonals are simultaneously frustrated. Specifically, the training data are sampled from a coupling distribution $\mathcal{P}(\boldsymbol{\gamma})$ defined exclusively on points of the form $(J_{2L}/J_1, 0)$ or $(0, J_{2R}/J_1)$, where only one of the two next-nearest-neighbor couplings is active, with J_{2L} and J_{2R} harvested from a uniform distribution defined on the interval $[0.0, 0.6]$. We then assess whether the resulting model can generalize to coupling configurations with $J_{2L} = J_{2R}$, which recover the J_1–J_2 Heisenberg model on a square lattice [20, 41], where frustration is introduced symmetrically along both diagonals. Importantly, such test points lie outside the support of the training distribution $\mathcal{P}(\boldsymbol{\gamma})$, challenging the model's

ability to extrapolate beyond its training regime. In Fig. 7.6b, we consider a 6×6 lattice and plot the spin–spin correlation function at two in-distribution points, $(J_{2L}/J_1, J_{2R}/J_1) = (0.0, 0.3)$ and $(0.3, 0.0)$, as well as at the out-of-distribution point $(0.3, 0.3)$. Remarkably, the FNQS accurately captures the enhanced frustration in the latter case, producing correlation functions that have lower amplitudes than in the case with only one frustrated diagonal at a time, and in close agreement with exact results. This demonstrates the model's surprising ability to generalize beyond the support of the training distribution. However, it is important to emphasize that the accuracy of such generalization decreases as the distance from the training axis ($J_{2L} = 0$ or $J_{2R} = 0$) increases. Specifically, along the diagonal direction ($J_{2L} = J_{2R}$), the relative error on the ground-state energy remains very small in the Néel antiferromagnetic phase of the J_1–J_2 Heisenberg model, on the order of 10^{-5} for $J_2 = 0.1$, 10^{-4} for $J_2 = 0.3$, but increases substantially to order 10^{-2} for $J_2 = 0.5$ and 10^{-1} for $J_2 = 0.6$. This degradation in generalization performance is illustrated in Fig. 7.6c, which shows the behavior of the Néel antiferromagnetic order parameter $m^2 = S(\pi, \pi)/N$, where $S(\boldsymbol{k}) = \sum_{\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \langle \hat{\boldsymbol{S}}_{\boldsymbol{0}} \cdot \hat{\boldsymbol{S}}_{\boldsymbol{r}} \rangle$ is the spin structure factor and N is the total number of sites in the J_1–J_2 Heisenberg model ($J_{2L} = J_{2R}$).

7.4 Discussion

In this chapter, we have demonstrated that a single neural-network architecture can be efficiently trained on multiple many-body quantum systems, yielding a variational state that generalizes to previously unseen coupling parameters. This approach enables the use of pre-trained states as starting points for specific investigations [141], similar to current practices in machine learning. To facilitate the adoption of this methodology, we have made FNQS models available through the Hugging Face Hub at `https://huggingface.co/nqs-models`, integrated with the `transformers` library [259] and providing simple interfaces for `NetKet` [145].

Future perspectives

Factored attention mechanism

The results on the factored attention layer in Chapter 1 highlight the potential of factored attention as a component of deep transformer models. In the future, it will be intriguing to extend the approach developed in Chapter 2 for constructing clones and for analyzing the learning dynamics of transformers to different data modalities, like protein sequences. On the theoretical side, developing a quantitative treatment of the sequential learning dynamics of (deep) transformers looms as an intriguing open challenge.

Scaling the ViT wavefunction to bigger system sizes

In the future, an important contribution to the field of Neural Quantum States will be to extend the techniques we have developed to study larger 2D systems, for example reaching sizes such as 30×30 or even 40×40. Having accurate descriptions of large systems will be crucial for solving many long-standing debates in physics. This will require adapting the Transformer architecture, with which we have currently achieved accurate results up to size 18×18 (see Chapter 5), but whose computational complexity scales quadratically with input size. For instance, one could use an architecture similar to the SwinTransformer [260]. In existing transformer-based models, including the one we have used to parametrize wave functions, the inputs are all of a fixed scale, a property unsuitable for vision applications and to describe correlations of certain quantum states. SwinTransformer constructs a hierarchical representation of the inputs by starting from small-sized patches and gradually merging neighboring patches in deeper Transformer layers. Moreover, SwinTransformer implements the shifted-window approach, which reduces the computational complexity to linear in the input size.

ViT wavefunction for fermionic systems

Another interesting and so far unexplored research direction is using the transformer architectures proposed in this Thesis for studying fermionic systems, for

example the Hubbard model on the square lattice. The latter is the most "simple" model of an interacting quantum system, but despite its simplicity and many intense numerical and theoretical studies, a full computational solution is still to be achieved [261, 262]. One of the central open questions concerns the nature of the ground state in the underdoped regime, where it is still debated whether the phase diagram hosts a genuine superconducting phase, stripe order, or some other form of intertwined correlations [263, 264]. The ViT wavefunction provides a promising alternative to traditional methods, as, in some cases we described in the Thesis, it provides a powerful ansatz capable of capturing complex spatial correlations and scaling effectively with system size (see Chapter 4). Future efforts should focus on adapting such architectures to fermionic systems and assessing their capacity to resolve long-standing open questions in the Hubbard model, including the presence or absence of superconducting and stripe-ordered phases. To study fermionic systems, the transformer could be used to build a backflow transformation following Ref. [265].

Foundation models for Neural Quantum States

In Chapter 7, we have restricted ourselves to the case where the different spin systems under study are obtained by fixing a single Hamiltonian and varying its coupling parameters. However, the developed framework is general and can be employed to simultaneously optimize the ground state of Hamiltonians defined on lattices with different geometries, or lattices with a different number of sites. These could represent interesting research directions to explore in the future, going in the promising direction of building a single FNQS capable of treating all spin models with arbitrary two-body interactions in two dimensions. Achieving this ambitious goal will require a step-by-step approach and will require generating wave functions for several models and systems, plausibly within a long-term research program. Moreover, the extension to fermionic systems in second quantization [266, 267] requires adapting the architecture while maintaining the core methodology. For molecular systems [268], the multimodal structure of FNQS could enable efficient computation of energy derivatives with respect to geometric parameters, providing access to atomic forces and equilibrium configurations. Beyond ground states, these foundation models could potentially facilitate the study of quantum dynamics by introducing explicit time-dependent variational states [269, 270], particularly in large systems where traditional methods become intractable. These developments, combined with the public availability of pretrained models, represent a step toward making advanced quantum many-body techniques more accessible to the broader physics community.

Neural Quantum States interpretability

A so far unexplored research direction is the study of the representations created by the Transformer during the optimization of the ground state (see Chapter 4). Understanding how these representations evolve and what physical information they encode could provide valuable insights into the inductive biases of NQS. A first step toward this goal could involve analyzing the geometry of the learned representations using unsupervised learning techniques, such as the BID framework introduced in Ref. [271], which has shown success in uncovering latent structure in Transformers trained on natural language tasks [272]. An ambitious goal would be to detect phase transitions directly from the feature space, as suggested in Ref. [273]. Such an approach could offer a model-agnostic route to identifying critical behavior without relying on prior knowledge of the relevant order parameters.

Acknowledgements

Ringrazio i miei supervisor, Alessandro e Sebastian, per gli insegnamenti ricevuti. Ringrazio Luciano per tutto il tempo trascorso insieme. Ringrazio la mia famiglia per il supporto. Ringrazio le persone che ho conosciuto a Trieste, per aver reso questa esperienza indimenticabile.

Chapter 8

Appendix

8.1 Appendix to Chapter 1

8.1.1 Replica analysis of self-attention

In this appendix we discuss in detail the replica analysis of a factored single layer self-attention. The computation builds upon the recent advances in statistical physics of learning, concerning the extension of replica theory to structured data [61, 63, 72]. In the following we will show how to slightly modify this new approach in order to deal with masked language modeling tasks, under the following simplified assumptions: mean-field limit, full-rank J matrix, U matrix fixed to the identity and thus not learned.

8.1.2 Statistical physics formulation of machine learning problems

Statistical physics considers learning as a dynamical and exploratory process across the space of the learnable parameters. At equilibrium, these parameters are assumed to follow a Boltzmann-Gibbs distribution, where the role of the Hamiltonian is actually played by the loss function:

$$\begin{aligned} \pi_\beta\left(\mathbf{A}_i, \mathcal{D}\right) &= \frac{P\left(\mathbf{A}_i\right)}{\mathcal{Z}_\beta} e^{-\beta \sum_{\mu=1}^{M} \ell\left(m_i^\mu, \frac{\mathbf{A}_i \cdot \mathbf{m}_{\setminus i}^\mu}{\sqrt{L}}\right)} \\ &= \frac{P_A\left(\mathbf{A}_i\right)}{\mathcal{Z}_\beta} \prod_{\mu=1}^{M} P_G\left(m_i^\mu | \frac{\mathbf{A}_i \cdot \mathbf{m}_{\setminus i}^\mu}{\sqrt{L}}\right) \end{aligned} \tag{8.1}$$

with β being the inverse temperature and $\mathcal{D}$ the training set. In the zero temperature limit (i.e. $\beta \to \infty$), the Boltzmann-Gibbs distribution concentrates around

the minima of the loss function, which are merely the solutions of the optimization problem in eq. 1.6:

$$\hat{\mathbf{A}}_i \underset{\beta\to\infty}{=} \mathbb{E}_{\pi_\beta}\left[\mathbf{A}_i\right] \tag{8.2}$$

Up to this point, re-framing a machine learning problem in terms of statistical physics did not seem to be very advantageous since sampling from a high-dimensional Boltzmann-Gibbs distribution is known to be impracticable.

This is where the replica theory comes into play. In particular, it states that, in the high-dimensional limit (i.e. $M, L \to \infty$ with $\alpha \equiv M/L \sim O(1)$) the free-energy of a learning system concentrates around its typical value over the input data distribution:

$$f_\beta = -\lim_{L\to\infty} \frac{1}{L} \mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}} \left[\log \mathcal{Z}_\beta\right]. \tag{8.3}$$

As we will see in the next section, this expectation can be tackled by means of the replica trick. From this quantity, all the high-dimensional metrics of interests, can be computed as a function of simple scalar quantities. This is for instance the case of the generalization loss in eq. (1.7). In particular, the overlap parameters $m_\star$ and $q_\star$ correspond to practically measurable quantities over different realisation of the training set, involving the estimator of the ith row of the interaction matrix:

$$q^\star = \frac{\hat{\mathbf{A}}_i^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{L} \qquad r^\star = -\frac{\mathbf{J}_i^t \Sigma_{\setminus i} \hat{\mathbf{A}}_i}{\nu L}. \tag{8.4}$$

In the next section we will outline the main steps of the replica trick leading to the generalization loss formula in eq. (1.7).

8.1.3 Replica Calculation

As anticipated in the previous section, the replica trick allows to compute the typical value of the free-energy density in eq. (8.3) by expressing this quantity as a function of the solely replicated partition function $\mathcal{Z}_\beta^n$, obtained by constructing $n > 0$ different and independent copies of the same learning system:

$$f_\beta = -\lim_{n\to 0^+} \frac{d}{dn} \lim_{L\to\infty} \left[\frac{1}{L} \mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i}, m^\mu_i\}} \mathcal{Z}_\beta^n\right]. \tag{8.5}$$

Average over the training set. As a first step, the replica calculation focuses on the expectation of the replicated partition function over the training set, which,

written in a more explicit form, looks like:

$$\begin{aligned}\mathbb{E}_{\{\mathbf{m}^{\mu}_{\setminus i}, m^{\mu}_{i}\}}\mathcal{Z}^{n}_{\beta} = &\int \prod_{a=1}^{n} d\mathbf{A}^{a}_{i} \prod_{a=1}^{n} P_{A}\left(\mathbf{A}^{a}_{i}\right) \\ &\times \prod_{\mu=1}^{M} \mathbb{E}_{m^{\mu}_{i}|\mathbf{m}^{\mu}_{\setminus i}} \mathbb{E}_{\mathbf{m}^{\mu}_{\setminus i}} \left[P_{G}\left(m^{\mu}_{i} \mid \frac{\mathbf{A}^{a}_{i} \cdot \mathbf{m}^{\mu}_{\setminus i}}{\sqrt{L}} \right) \right],\end{aligned} \tag{8.6}$$

with P_G and P_A being respectively the Gibbs and the Gaussian measure associated with the ith row of the attention matrix as in eq. (8.1). Indeed, as already pointed out in the main manuscript, the interaction matrix is drawn from the Gaussian Orthogonal Ensemble, therefore its rows will correspond to Gaussian random vectors. At this point, we can notice an important aspect of MLM tasks. In this case, the labels are not provided by a teacher vector as in standard teacher-student settings. On the contrary, the masked tokens are directly sampled from the input distribution by conditioning over all the other elements composing the sequence:

$$\begin{aligned} m^{\mu}_{i} &\sim \frac{P_{J}\left(\mathbf{J}_{i}\right)}{\sqrt{2\pi\nu^{-1}}} e^{-\frac{1}{2\nu^{-1}}\left(m^{\mu}_{i} + \frac{\mathbf{J}_{i}\cdot\mathbf{m}^{\mu}_{\setminus i}}{\nu\sqrt{L}}\right)^{2}} \\ &= P_{J}\left(\mathbf{J}_{i}\right) P_{0}\left(m^{\mu}_{i} \mid \frac{\mathbf{J}_{i} \cdot \mathbf{m}^{\mu}_{\setminus i}}{\nu\sqrt{L}} \right). \end{aligned} \tag{8.7}$$

Note that, the noise in the labels arises a consequence of the one already affecting the input data points, meaning that its intensity can not be chosen independently from the intrinsic stochasticity of the input. Due to these considerations, by explicitly expressing the outer expectation in eq. (8.6), we then obtain

$$\begin{aligned}\mathbb{E}_{\{\mathbf{m}^{\mu}_{\setminus i}, m^{\mu}_{i}\}}\mathcal{Z}^{n}_{\beta} = &\int d\mathbf{J}_{i}\ P_{J}\left(\mathbf{J}_{i}\right) \prod_{a=1}^{n} \left(\int d\mathbf{A}^{a}_{i}\ P_{A}\left(\mathbf{A}^{a}_{i}\right) \right) \times \\ &\times \prod_{\mu=1}^{M} \left(\int dm^{\mu}_{i} \mathbb{E}_{\mathbf{m}^{\mu}_{\setminus i}} \left[P_{0}\left(m^{\mu}_{i} \mid \frac{\mathbf{J}_{i} \cdot \mathbf{m}^{\mu}_{\setminus i}}{\nu\sqrt{L}} \right) \times \right.\right. \\ &\left.\left. \times P_{G}\left(m^{\mu}_{i} \mid \frac{\mathbf{A}^{a}_{i} \cdot \mathbf{m}^{\mu}_{\setminus i}}{\sqrt{L}} \right) \right] \right).\end{aligned} \tag{8.8}$$

To compute the expectation over the input sequences with a masked element at position i, we first define the pre-activations as:

$$h^{\mu}_{a} = \frac{\mathbf{A}^{a}_{i} \cdot \mathbf{m}^{\mu}_{\setminus i}}{\sqrt{L}} \qquad z^{\mu} = -\frac{\mathbf{J}_{i} \cdot \mathbf{m}^{\mu}_{\setminus i}}{\nu\sqrt{L}}, \tag{8.9}$$

then we express these definitions in terms of Dirac-deltas and their corresponding integral representation:

$$
\begin{aligned}
1 &\propto \int \prod_{\mu=1}^{M} \prod_{a=1}^{n} \frac{dh_a^\mu d\hat{h}_a^\mu}{2\pi} \prod_{\mu=1}^{M} e^{i\hat{h}_a^\mu \left(h_a^\mu - \frac{\mathbf{A}_i^a \cdot \mathbf{m}_{\setminus i}^\mu}{\sqrt{L}} \right)} \\
1 &\propto \int \prod_{\mu=1}^{M} \frac{dz^\mu d\hat{z}^\mu}{2\pi} \prod_{\mu=1}^{M} e^{i\hat{z}^\mu \left(z^\mu + \frac{\mathbf{J}_i \cdot \mathbf{m}_{\setminus i}^\mu}{\nu\sqrt{L}} \right)}.
\end{aligned}
\tag{8.10}
$$

By plugging these factors one into eq. (8.8), we then get:

$$
\begin{aligned}
\mathbb{E}_{\{\mathbf{m}_{\setminus i}^\mu, m_i^\mu\}} \mathcal{Z}_\beta^n &= \int d\mathbf{J}_i \; P_J(\mathbf{J}_i) \int \prod_{a=1}^{n} d\mathbf{A}_i^a \; \prod_{a=1}^{n} P_A(\mathbf{A}_i^a) \times \\
&\times \int \prod_{\mu=1}^{M} dm_i^\mu \int \prod_{\mu=1}^{M} \frac{dz^\mu d\hat{z}^\mu}{2\pi} \prod_{\mu=1}^{M} e^{i\hat{z}^\mu z^\mu} \prod_{\mu=1}^{M} P_0(m_i^\mu \mid z^\mu) \times \\
&\times \int \prod_{\mu=1}^{M} \prod_{a=1}^{n} \frac{dh_a^\mu d\hat{h}_a^\mu}{2\pi} \prod_{\mu=1}^{M} \prod_{a=1}^{n} e^{i\hat{h}_a^\mu h_a^\mu} \prod_{\mu=1}^{M} \prod_{a=1}^{n} P_G(m_i^\mu \mid h_a^\mu) \times \\
&\times \prod_{\mu=1}^{n} \mathbb{E}_{\mathbf{m}_{\setminus i}^\mu} \left[e^{i\hat{z}^\mu \frac{\mathbf{J}_i \cdot \mathbf{m}_{\setminus i}^\mu}{\nu\sqrt{L}}} \prod_{a=1}^{n} e^{-i\hat{h}_a^\mu \frac{\mathbf{A}_i^a \cdot \mathbf{m}_{\setminus i}^\mu}{\sqrt{L}}} \right].
\end{aligned}
\tag{8.11}
$$

The expectation over the masked input sequences is a simple multivariate Gaussian integral, whose solution is given by:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{m}_{\setminus i}^\mu} \left[e^{i\hat{z}^\mu \frac{\mathbf{J}_i \cdot \mathbf{m}_{\setminus i}^\mu}{\nu\sqrt{L}}} \prod_{a=1}^{n} e^{-i\hat{h}_a^\mu \frac{\mathbf{A}_i^a \cdot \mathbf{m}_{\setminus i}^\mu}{\sqrt{L}}} \right] &= e^{-\frac{1}{2} \frac{\mathbf{J}_i^t \Sigma_{\setminus i} \mathbf{J}_i}{\nu^2 L} (z^\mu)^2} \times \\
\times \, e^{\sum_{a=1}^{n} \frac{\mathbf{J}_i^t \Sigma_{\setminus i} \mathbf{A}_i^a}{\nu L} \hat{h}_a^\mu \hat{z}^\mu - \frac{1}{2} \sum_{a,b=1}^{n} \frac{(\mathbf{A}_i^a)^t \Sigma_{\setminus i} \mathbf{A}_i^a}{L} \hat{h}_a^\mu \hat{h}_b^\mu}
\end{aligned}
\tag{8.12}
$$

where, as already pointed out in the main text, $\Sigma_{\setminus i}$ corresponds to the covariance matrix of the masked input sequences, that is the input covariance matrix without the contribution of the row and column associated to the ith masked token. By

replacing the solution of the expectation in eq. (8.11), we then get:

$$\begin{aligned}\mathbb{E}_{\{\mathbf{m}^{\mu}_{\setminus i},m^{\mu}_{i}\}}\mathcal{Z}^{n}_{\beta} &= \int d\mathbf{J}_i\ P_J\left(\mathbf{J}_i\right)\int\prod_{a=1}^{n}d\mathbf{A}^{a}_{i}\ \prod_{a=1}^{n}P_A\left(\mathbf{A}^{a}_{i}\right)\times\\ &\times\int\prod_{\mu=1}^{M}dm^{\mu}_{i}\int\prod_{\mu=1}^{M}\frac{dz^{\mu}d\hat{z}^{\mu}}{2\pi}\prod_{\mu=1}^{M}e^{i\hat{z}^{\mu}z^{\mu}}\prod_{\mu=1}^{M}P_0\left(m^{\mu}_{i}\mid z^{\mu}\right)\times\\ &\times\int\prod_{\mu=1}^{M}\prod_{a=1}^{n}\frac{dh^{\mu}_{a}d\hat{h}^{\mu}_{a}}{2\pi}\prod_{\mu=1}^{M}\prod_{a=1}^{n}e^{i\hat{h}^{\mu}_{a}h^{\mu}_{a}}\prod_{\mu=1}^{M}\prod_{a=1}^{n}P_G\left(m^{\mu}_{i}\mid h^{\mu}_{a}\right)\times\\ &\times\prod_{\mu=1}^{n}e^{-\frac{1}{2}\frac{\mathbf{J}^{t}_{i}\Sigma_{\setminus i}\mathbf{J}_i}{\nu^2L}(z^{\mu})^2+\sum_{a=1}^{n}\frac{\mathbf{J}^{t}_{i}\Sigma_{\setminus i}\mathbf{A}^{a}_{i}}{\nu L}\hat{h}^{\mu}_{a}\hat{z}^{\mu}-\frac{1}{2}\sum_{a,b=1}^{n}\frac{\left(\mathbf{A}^{a}_{i}\right)^{t}\Sigma_{\setminus i}\mathbf{A}^{a}_{i}}{L}\hat{h}^{\mu}_{a}\hat{h}^{\mu}_{b}}.\end{aligned}\tag{8.13}$$

Rewriting the averaged replicated partition function in terms of saddle-point integrals. As a consequence of the average over the training set, the different replicas are now interacting among each other through the following set of overlap parameters:

$$\rho=\frac{\mathbf{J}^{t}_{i}\Sigma_{\setminus i}\mathbf{J}_i}{\nu^2L}\qquad r_a=-\frac{\left(\mathbf{A}^{a}_{i}\right)^{t}\Sigma_{\setminus i}\mathbf{J}_i}{\nu L}\qquad q_{ab}=-\frac{\left(\mathbf{A}^{a}_{i}\right)^{t}\Sigma_{\setminus i}\mathbf{A}^{b}_{i}}{L}.\tag{8.14}$$

Once again, to proceed further in the calculation, we can insert their definition by means of Dirac-deltas and their integral representation:

$$\begin{aligned}1&\propto\int\frac{d\rho d\hat{\rho}}{2\pi}\ e^{i\hat{\rho}\left(\nu^2L\rho-\mathbf{J}^{t}_{i}\Sigma_{\setminus i}\mathbf{J}_i\right)}\\ 1&\propto\int\prod_{a=1}^{n}\frac{dr_a d\hat{r}_a}{2\pi}\prod_{a=1}^{n}\ e^{i\hat{r}_a\left(-\nu Lr_a-\left(\mathbf{A}^{a}_{i}\right)^{t}\Sigma_{\setminus i}\mathbf{J}_i\right)}\\ 1&\propto\int\prod_{a\leq b}\frac{dq_{ab}d\hat{q}_{ab}}{2\pi}\prod_{a\leq b}\ e^{i\hat{q}_{ab}\left(q_{ab}-\left(\mathbf{A}^{a}_{i}\right)^{t}\Sigma_{\setminus i}\mathbf{A}^{b}_{i}\right)}.\end{aligned}\tag{8.15}$$

By substituting the overlap definition in eq. (8.13), plugging in the corresponding factors one and performing the change of variables: $i\hat{\rho}\to-\hat{\rho}$, $i\hat{r}_a\to\hat{r}_a$ and $i\hat{q}_{ab}\to\hat{q}_{ab}$, we can rewrite the averaged replicated partition function in terms of saddle-point integrals over the overlap parameters:

$$\mathbb{E}_{\{\mathbf{m}^{\mu}_{\setminus i},m^{\mu}_{i}\}}\mathcal{Z}^{n}_{\beta}=\int\frac{d\rho d\hat{\rho}}{2\pi}\int\prod_{a=1}^{n}\frac{dr_a d\hat{r}_a}{2\pi}\int\prod_{a\leq b}\frac{dq_{ab}d\hat{q}_{ab}}{2\pi}\ e^{L\Psi^{(n)}}\tag{8.16}$$

where the action $\Psi^{(n)}$ is a non trivial function of the overlap parameters:

$$\Psi^{(n)}=-\nu^2\rho\hat{\rho}+\nu\sum_{a=1}^{n}r_a\hat{r}_a-\sum_{a\leq b}q_{ab}\hat{q}_{ab}+\frac{1}{L}\Psi_s+\frac{M}{L}\Psi_e\tag{8.17}$$

where Ψ_s and Ψ_e are the so-called entropic and energetic potential and, in the specific case of a single-layer factored attention, are given by:

$$\begin{aligned}
\Psi_s &= \log\Bigg[\int d\mathbf{J}_i P_J\left(\mathbf{J}_i\right)\int\prod_{a=1}^{n}\left(d\mathbf{A}_i^a P_A\left(\mathbf{A}_i^a\right)\right)\times \\
&\quad\times e^{\hat{\rho}\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i+\sum_{a=1}^{n}\hat{r}_a\left(\mathbf{A}_i^a\right)^t\Sigma_{\setminus i}\mathbf{J}_i+\sum_{a\leq b}\hat{q}_{ab}\left(\mathbf{A}_i^a\right)^t\Sigma_{\setminus i}\mathbf{A}_i^b}\Bigg] \\
\Psi_e &= \log\Bigg[\int dm_i\int\frac{dzd\hat{z}}{2\pi}\;e^{-\frac{\rho}{2}\hat{z}^2+i\hat{z}z}P_0\left(m_i|z\right)\times \\
&\quad\times\int\prod_{a=1}^{n}\frac{dh_a d\hat{h}_a}{2\pi}\;e^{-\frac{1}{2}\sum_{a,b=1}^{n}q_{ab}\hat{h}_a\hat{h}_b-\hat{z}\sum_{a=1}^{n}r_a\hat{h}_a}\times \\
&\quad\times\; e^{i\sum_{a=1}^{n}\hat{h}_a h_a}P_G\left(m_i|h_a\right)\Bigg].
\end{aligned} \tag{8.18}$$

Note that, we have drop the dependency of Ψ_e on the μ index since all the μ-dependent terms decouple with respect to μ.

Replica Symmetric Assumption. To proceed further in the calculation, we need to assume a specific replica structure. Since all replicas have been introduced independently from each other with no specific differences among them, it seems natural to assume that replicas should all play the same role and that, therefore, the overlap parameters should not depend on the specific replica index. In particular, under the replica symmetric ansatz, we assume:

$$\begin{aligned}
q_{ab} &= \begin{cases} g & \text{if } a=b \\ q & \text{otherwise} \end{cases} \qquad -i\hat{q}_{ab} = \begin{cases} -\frac{\hat{g}}{2} & \text{if } a=b \\ \hat{q} & \text{otherwise} \end{cases} \\
r_a &= r \qquad\qquad -i\hat{r}_a = \hat{r} \quad \forall a
\end{aligned} \tag{8.19}$$

By plugging the replica symmetric assumption in eq. (8.17)-(8.18) and applying the following Hubbard Stratonovich transformations:

$$\begin{aligned}
e^{\frac{\hat{q}}{2}\sum_{a\leq b}(\mathbf{A}^a)^t\Sigma_{\setminus i}\mathbf{A}^b} &= \int\mathcal{D}\boldsymbol{\xi}\;\exp\left(\sum_{a=1}^{n}\left(\mathbf{A}_i^a\right)^t\left(\hat{q}\Sigma_{\setminus i}\right)^{1/2}\boldsymbol{\xi}\right) \\
e^{-\frac{q}{2}\sum_{a,b=1}^{n}\hat{h}_a\hat{h}_b} &= \int\mathcal{D}\xi\;\exp\left(i\sqrt{q}\sum_{a=1}^{n}\hat{h}_a^2\;\xi\right),
\end{aligned} \tag{8.20}$$

with $\xi \sim \mathcal{N}(0,1)$, we then get the expression for the replica symmetric action:

$$\begin{aligned}\Psi^{(n)} &= -\nu^2\rho\hat{\rho} + \nu r\hat{r} + \frac{n}{2}(\delta q + q)(\delta\hat{q} - \hat{q}) - \frac{n(n-1)}{2}q\hat{q} \\ &+ \frac{1}{L}\Psi_s + \frac{M}{L}\Psi_e\end{aligned} \tag{8.21}$$

where we have defined $\delta\hat{q} = \hat{g} + \hat{q}$ and $\delta q = g - q$. the replica symmetric potentials Ψ_s and Ψ_e are are given by:

$$\begin{aligned}\Psi_s &= \log\left[\int d\mathbf{J}_i P_J(\mathbf{J}_i)\, e^{\hat{\rho}\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i}\int \mathcal{D}\boldsymbol{\xi}\times\right. \\ &\left.\times\left(\int d\mathbf{A}_i P_A(\mathbf{A}_i)\, e^{\hat{r}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{J}_i - \frac{\hat{l}}{2}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{A}_i + \mathbf{A}_i^t\left(\hat{q}\Sigma_{\setminus i}\right)^{1/2}\boldsymbol{\xi}}\right)^n\right] \\ \Psi_e &= \log\left[\int dm_i \int \frac{dzd\hat{z}}{2\pi}\; e^{-\frac{\rho}{2}\hat{z}^2 + i\hat{z}z} P_0(m_i|z)\int \mathcal{D}\xi\times\right. \\ &\left.\times\left(\int \frac{dhd\hat{h}}{2\pi}\; e^{-\frac{l}{2}\hat{h}^2 + i\sqrt{q}\xi\hat{h} - m\hat{z}\hat{h} + i\hat{h}h} P_G(m_i|h)\right)^n\right]\end{aligned} \tag{8.22}$$

Zero Replica limit. By taking the limit of $n \to 0$ in eq. (8.21)-(8.22) and solving the integrals with respect to the $\hat{z}$ and $\hat{h}$ variables, we then get the following expression for the action potential:

$$\Psi^{(n\to 0)} = \nu r\hat{r} + \frac{1}{2}(l+q)\left(\hat{l} - \hat{q}\right) + \frac{1}{2}q\hat{q} + \frac{1}{L}\Psi_s^{(n\to 0)} + \frac{M}{L}\Psi_e^{(n\to 0)} \tag{8.23}$$

with the entropic and energetic potentials in the zero replicas limit given by:

$$\begin{aligned}\Psi_s^{(n\to 0)} &= \int \mathcal{D}\boldsymbol{\xi}\int d\mathbf{J}_i P_J(\mathbf{J}_i)\times \\ &\log\left[\int d\mathbf{A}_i P_A(\mathbf{A}_i)\, e^{\hat{r}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{J}_i - \frac{\hat{l}}{2}\mathbf{A}_i^t\Sigma_{\setminus i}\mathbf{A}_i + \mathbf{A}_i^t\left(\hat{q}\Sigma_{\setminus i}\right)^{1/2}\boldsymbol{\xi}}\right] \\ \Psi_e^{(n\to 0)} &= \int \mathcal{D}\xi\int dm_i \int \frac{dz}{\sqrt{2\pi\left(\rho - \frac{r^2}{q}\right)}}\; e^{-\frac{\left(z + \frac{r}{\sqrt{q}}\xi\right)^2}{2\left(\rho - \frac{r^2}{q}\right)}}\times \\ &P_0(m_i|z)\log\left[\int \frac{dh}{\sqrt{2\pi l}}\; e^{-\frac{(h+\sqrt{q}\xi)^2}{2l}} P_G(m_i|h)\right]\end{aligned} \tag{8.24}$$

Note that, as in standard teacher-student settings [61], in order to avoid divergent terms in this limit, the overlap ρ and its conjugate $\hat{\rho}$ need to be constrained to $\mathbb{E}_{\mathbf{J}_i}\left[\mathbf{J}_i^t\Sigma_{\setminus i}\mathbf{J}_i\right]/\nu^2$ and 0 respectively.

Typical Free-Energy density. Having determined the expression for the replicated partition function in the zero-temperature limit, we can actually compute the typical free-energy density as:

$$\begin{aligned} f_\beta &= -\lim_{n\to 0^+}\frac{d}{dn}\lim_{L\to\infty}\left[\frac{1}{L}\mathbb{E}_{\{\mathbf{m}^\mu_{\setminus i},m^\mu_i\}}\mathcal{Z}^n_\beta\right] \\ &= \lim_{L\to\infty}\frac{1}{L}\int\frac{d\rho d\hat{\rho}}{2\pi}\int\frac{drd\hat{r}}{2\pi}\int\frac{dqd\hat{q}}{2\pi}\; e^{L\Psi^{(n\to 0)}}. \end{aligned} \tag{8.25}$$

In the high-dimensional limit, we can solve the integrals over the overlap parameters by saddle-point, thus getting:

$$\begin{aligned} f_\beta = \underset{q,r,\delta q,\hat{q},\hat{r},\delta\hat{q}}{\text{extr}}\Bigg[&\nu r\hat{r}+\frac{1}{2}\left(\delta q+q\right)\left(\delta\hat{q}-\hat{q}\right)+\frac{1}{2}q\hat{q}+ \\ &+\lim_{L\to\infty}\frac{1}{L}\Psi_s^{(n\to 0)}+\alpha\Psi_e^{(n\to 0)}\Bigg], \end{aligned} \tag{8.26}$$

where the values of the overlap parameters extremizing the typical free-energy density can therefore be determined by solving the following system of coupled saddle-point equations:

$$\begin{cases} q=-2\frac{\partial\Psi_s^{(n\to 0)}}{\partial\delta\hat{q}} \\ \delta q=2\frac{\partial\Psi_s^{(n\to 0)}}{\partial\hat{q}} \\ r=-\frac{1}{\nu}\frac{\partial\Psi_s^{(n\to 0)}}{\partial\hat{r}} \end{cases} \qquad \begin{cases} \hat{q}=2\frac{\partial\Psi_e^{(n\to 0)}}{\partial\delta q} \\ \delta\hat{q}=-2\frac{\partial\Psi_e^{(n\to 0)}}{\partial q} \\ \hat{r}=-\frac{1}{\nu}\frac{\partial\Psi_e^{(n\to 0)}}{\partial r} \end{cases}. \tag{8.27}$$

Up to this point, we have performed the replica calculation in full generality, without specifying neither the interaction matrix nor the loss function. In the next section, we will evaluate the typical free-energy density for the specific MLM task of eq. (1.6) under the simplified assumptions of the sec. *The sample complexity of self-attention* of the main text.

Zero temperature limit and Gaussian Priors. As already pointed out in the main text, the interaction matrix J is sampled from the GOE, it is then natural to assume a Gaussian prior on the ith row of the attention matrix:

$$P_A\left(\mathbf{A}\right)=\frac{1}{\sqrt{2\pi}}e^{\beta\lambda\mathbf{A}_i^t\mathbf{A}_i} \tag{8.28}$$

with β being the inverse temperature parameter, while λ is the L_2 regularization strength. Moreover, the optimization problem in eq. (1.6) optimizes a square loss

to solve the corresponding MLM task. This means that, the Gibbs Measure of eq. (8.1) associated with this task is:

$$P_G\left(m_i|h\right) \propto e^{-\frac{\beta}{2}\left(m_i-h\right)^2}. \tag{8.29}$$

By plugging these two specific forms of both the prior and the Gibbs measure in eq. (8.6) and taking the zero temperature limit as exemplified in [61, 72], we get the following expression for the typical free-energy density in the zero-temperature limit:

$$\begin{aligned} f \underset{\beta\to\infty}{=} \underset{q,r,\delta q,\hat{q},\hat{r},\delta\hat{q}}{\text{extr}} \Bigg[& \nu r\hat{r} - \frac{1}{2}\left(\delta q\hat{q} - q\delta\hat{q}\right) + \\ & + \lim_{L\to\infty}\frac{1}{L}\Psi_s^{(n\to 0)} + \alpha\Psi_e^{(n\to 0)}\Bigg], \end{aligned} \tag{8.30}$$

with the entropic and energetic potential given by:

$$\begin{aligned} \Psi_s^{n\to 0} \underset{\beta\to\infty}{=} & \frac{1}{2}\Big[\hat{r}^2 \text{tr}\left(\Sigma_{\setminus i}\right)^t \Sigma_{\setminus i}\left(\beta\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i}\right)^{-1} + \\ & + \hat{q}\text{tr}\Sigma_{\setminus i}\left(\beta\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i}\right)^{-1}\Big] \\ \Psi_e^{n\to 0} \underset{\beta\to\infty}{=} & -\frac{1}{2}\frac{\nu^{-1} + \rho + q - 2r}{1+\delta q} \end{aligned} \tag{8.31}$$

and $\rho = \text{tr}\Sigma_{\setminus i}/\left(\nu^2 L\right)$. Note that, this functional form of the typical free-energy density corresponds exactly to the one of supervised learning with noisy label and Gaussian structured data [72]. However, we should point out again that, the variance of the noise in the labels, controlled by the shift factor ν, is a direct consequence of the intrinsic noise already affecting the input. Therefore, it can not be tuned independently from it. This is also reflected in the slightly different functional forms of the saddle-point equations in (8.27), which, in the zero-temperature limit with Gaussian priors and square loss, are given by:

$$\begin{cases} q = \text{tr}\left[\left(\hat{q}\Sigma_{\setminus i} + \hat{r}^2\Sigma_{\setminus i}\left(\Sigma_{\setminus i}\right)^t\right)\Sigma_{\setminus i}\left(\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i}\right)^{-2}\right] \\ \delta q = \text{tr}\left[\left(\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i}\right)^{-1}\Sigma_{\setminus i}\right] \\ r = -\frac{\hat{r}}{\nu}\text{tr}\left[\Sigma_{\setminus i}\left(\Sigma_{\setminus i}\right)^t\left(\lambda\mathbb{I} + \delta\hat{q}\Sigma_{\setminus i}\right)^{-1}\right] \\ \hat{q} = \frac{\nu^{-1}+\rho+q-2r}{\left(1+\delta q\right)^2} \\ \delta\hat{q} = \frac{1}{1+\delta q} \\ \hat{r} = -\frac{1}{\nu(1+\delta q)} \end{cases}. \tag{8.32}$$

As in the case of supervised learning settings [61, 72], the solution of this system of coupled saddle-point equations in the zero temperature limit allows us to express the generalization loss as shown in Eq. 1.7 of the main text, with the exception that the noise in the labels is the direct consequence of the intrinsic noise of the inputs.

8.2 Appendix to Chapter 2

8.2.1 Analytical details of the dynamics of a factored attention layer

Solving for $A^{(2)}$ in the case $A^{(1)}(t) = 0$

We aim to solve the dynamical equations presented in Eqs. 2.9 of the main text, specifically in the scenario where the first layer is skipped, which corresponds to setting $A^{(1)}(t) = 0$. Under this condition, the dynamics of the second layer are described by:

$$\dot{A}_{0j}^{(2)} = h_j - \left(CA_0^{(2)}\right)_j \ .$$

To solve this differential equation, we first write $C = U^T D U$ with D a diagonal matrix with diagonal elements D_j and $UU^T = I$ since the covariance matrix is symmetric. Thus we have

$$\dot{\tilde{A}}_{0j}^{(2)} + D_j \tilde{A}_{0j}^{(2)} = \tilde{h}_j \ ,$$

where we have defined $\tilde{A}_{0j}^{(2)} = UA_{0j}^{(2)}$ and $\tilde{h}_j = Uh_j$. In this new basis, the equations are decoupled. With the initial conditions $\tilde{A}_{0j}^{(2)}(t=0) = \tilde{A}_{0j}^{(2)}(0)$, the solution is:

$$\tilde{A}_{0j}^{(2)}(t) = \frac{\tilde{h}_j}{D_j} + \left(\tilde{A}_{0j}^{(2)}(0) - \frac{\tilde{h}_j}{D_j}\right) e^{-D_j t} \ .$$

It is easy to verify that, in the original basis, the stationary solution is $A_{0j}^{(2)}(t \to \infty) = (C^{-1}h)_j$ as discussed in the main text.

Loss lower bound

Consider the analytical setup presented in section 2.3 of the main text. The mean square error loss associated with the model in Eq. 2.7 is:

$$\mathcal{L}(A^{(1)}, A_0^{(2)}) = \left(x_0 - \sum_j A_{0j}^{(2)} x_j - \sum_{jkl} A_{0j}^{(2)} A_{jk}^{(1)} A_{jl}^{(1)} x_k x_l\right)^2$$

where $b_i = B_{0i}$. We want to show that, in the case of Gaussian distributed data, for any values of the parameters b_i it is bounded from below by the value it has in $A^{(1)} = 0$. Taking the expectation, we get:

$$\mathcal{L}(A^{(1)} \mid A_0^{(2)}) = f(A_0^{(2)}) + \sum_{jkl} \sum_{j'k'l'} A_{0j}^{(2)} A_{jk}^{(1)} A_{jl}^{(1)} b_{j'} A_{j'k'}^{(1)} A_{j'l'}^{(1)} E\left[x_{k'} x_{l'} x_k x_l\right]$$

$$= f(A_0^{(2)}) + \sum_{jj'} b_{j'} A_{0j}^{(2)} \sum_{klk'l'} A_{jk}^{(1)} A_{jl}^{(1)} A_{j'k'}^{(1)} A_{j'l'}^{(1)} \left(C_{kl} C_{k'l'} + C_{kk'} C_{ll'} + C_{kl'} C_{k'l}\right)$$

$$= f(A_0^{(2)}) + \sum_{jj'} b_{j'} A_{0j}^{(2)} \left(\Gamma_{jj} \Gamma_{j'j'} + 2\Gamma_{jj'}^2\right) = f(A_0^{(2)}) + \left(\sum_j A_{0j}^{(2)} \Gamma_{jj}\right)^2 + 2 \sum_{jj'} b_{j'} A_{0j}^{(2)} \Gamma_{jj'}^2 \ ,$$

where $f(A_0^{(2)})$ is a function of $A_0^{(2)}$ alone, and we defined $\Gamma = A^{(1)} C A^{(1)}$. For any possible choice of the vector $A_0^{(2)}$, we have:

$$\mathcal{L}(A^{(1)} \mid A_0^{(2)}) \geq \mathcal{L}(A^{(1)} = 0 \mid A_0^{(2)}) = f(A_0^{(2)}) \ .$$

Indeed, $\left(\sum_j A_{0j}^{(2)} \Gamma_{jj}\right)^2 \geq 0$ for all choices of $A^{(1)}$. Moreover, $\sum_{jj'} b_{j'} A_{0j}^{(2)} \Gamma_{jj'}^2 \geq 0$, since all the matrices $\Gamma_{jj'}^2$ are non-negative; therefore, all its eigenvalues are non-negative.

8.2.2 Experimental setup for simplicity bias in Transformers

TinyStories is a synthetic data set available on Huggingface [274] and generated by GPT−3.5 and GPT−4, which contains a collection of stories that typical 3 to 4-year-olds usually understand [78]. This data set was recently introduced to facilitate the development, analysis and research of LLMs. In particular, it can be used to train and evaluate LLMs much smaller than state-of-the-art models, yet still being able to produce fluent stories with consistent meaning, thus demonstrating reasoning capabilities. This has been shown even for models with one layer and below 10 million total parameters [78]. In the same spirit of Ref. [93], we use the *full-sentences* training scheme for MLM settings. Therefore, we first concatenate all the stories of the data set, tokenising the resulting text with the ByteLevelBPETokenizer (as implemented in Huggingface) and setting the vocabulary size to $|\mathbb{V}| = 10000$. We then extract sequences of tokens of length $L = 128$ [275]; notice that, in the full-sentences scheme, the sequences may contain sentences that are unfinished or that cross the boundaries of different stories. In the training phase, we adopt a strategy wherein, for each sequence, we randomly mask 15% of the tokens individually, as opposed to a simultaneous masking approach.

WikiText We also used the WikiText-103 data set [94], which is another standard NLP data set. Similar to tinystories, we tokenized using the ByteLevelBPETokenizer, setting the vocabulary size to $|\mathbb{V}| = 500$ and considering sequences of length $L = 64$ tokens.

Obtaining the clones

We exploit the generative model described in Section 2.2 to generate clones of the original TinyStories data set (see Section 8.2.2 for details on how we prepare the data set). The clones are meant to approximate the underlying TinyStories distribution with increasing fidelity. In particular, as already discussed, by stacking n layers of factored attention, we can generate copies of TinyStories whose maximum degree of interaction is $n+1$. To enhance training stability, we normalize the activations through Layer Normalization. Each layer of factored attention involves two matrices of parameters, $A \in R^{L\times L}$ and $V \in R^{|\mathbb{V}|\times|\mathbb{V}|}$. In our setup, the size of the vocabulary is $|\mathbb{V}| = 10000$, thus the matrix V would involve 100 millions of parameters. To reduce this number, we perform a *finite*-rank approximation of the value matrix, writing $V = AB$, with $A \in R^{|\mathbb{V}|\times r}$ and $B \in R^{r\times|\mathbb{V}|}$. In practice, we choose $r = 400$. In this setting, we train two, four and six layers of factored attention via MLM. These models can then be sampled to generate three distinct clones of the original TinyStories data set, with a different degree of interaction among tokens.

Sampling procedure

Auto-regressive architectures [104] are trained to learn the set of distributions $\{p(s_i|s_{<i})\}_{i=1}^{L}$ and can be efficiently sampled. These models are associated with a joint over the tokens through the chain rule. On the other hand, bidirectional architectures model are trained to learn the set of conditionals $\{p(s_i|s_{\neq i})\}_{i=1}^{L}$. One may be tempted to sample these models with a Gibbs sampler, however this can result in an *inconsistent* sampler, where the conditionals may not be associated to any valid joint distribution. To understand why this could be the case, consider two discrete random variables X and Y, taking respectively m and n different possible values. The conditional distributions are defined in terms of two matrices $A_{ij} = P(X = x_i|Y = y_j)$ and $B_{ij} = P(Y = y_j|X = x_i)$, subject to the constraints $\sum_i A_{ij} = 1$ for every j and $\sum_j B_{ij} = 1$ for every i. The conditionals A and B are compatible if and only if two vectors $\tau_i = P(X = x_i)$ and $\eta_j = P(Y = y_j)$, defining the marginals, exist such that $\tau_i B_{ij} = \eta_j A_{ij}$ for every i and j. The latter is a linear system, which does not always have a solution (see Ref. [276] for more details and for the generalisation of the problem to more than two random variables). A possible solution is proposed in Ref. [95], where the authors develop a

sampling scheme based on the Metropolis Hastings (MH) algorithm, starting from the evidence that BERT can be used to reliably score a set of sequences [277]. In this approach, for each given input sequence $(s_1, \ldots, s_L)$, its score is defined as $E(s_1, \ldots, s_L) = -\sum_{i=1}^{L} h_i \cdot s_i$, with h_i being the logits as defined in Section 2.2. Then, the MH sampler is run on the joint distribution defined as $p(s_1, \ldots, s_L) \propto \exp(-E(s_1, \ldots, s_L))$. Starting from an initial sequence $(s_1^{t=0}, \ldots, s_L^{t=0})$, at each step a local move is proposed, in which the i-th token is masked and replaced with $s_i^{t+1} \sim p_{\text{mlm}}(s_i | s_{\neq i}^t)$. The move is then accepted according to the standard MH criterion. Hence, in this framework, the conditional distributions obtained after training the transformer with MLM are used as the transition kernel of the MH algorithm. In order to generate the clones of TinyStories, we choose randomly 620 examples of the test set and evolve them for 20 MH sweeps (2560 steps).

Architecture and the training procedure

BERT We use a *standard* BERT-like transformer encoder [92, 93] with input-dependent attention weights (thus using queries and keys) and MLP after each self-attention layer. We use skip connections and put LayerNorm inside each residual block, thus adopting the successful Pre-LN scheme introduced in Ref. [278]. We choose the number of layers equal to 4, the number of heads equal to 16, embedding dimension equal to 768 and the size of the hidden layer of the MLPs equal to 1536. Similar hyper-parameters are also used for the models trained in Ref. [78]. We choose SGD for the optimizer, setting the batch size to 1024. We start with a learning rate of 0.03, annealing it with a cosine decay scheduler. We trained the model for 10 hours on eight A100 GPUs.

GPT For our experiments with next-token prediction, we trained a two-layer GPT-Neo model, as also investigated in the TinyStories paper [78]. We set the hyperparameters of the network to be: number of layers equal to 2, number of heads equal to 16 and embedding dimension equal to 768. We used the AdamW optimiser, with the same hyperparameters used in Ref. [78], namely learning rate equal to $5e-4$, exponential decay $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The weight decay was set to 0.1 and we used a batch size equal to 128.

8.3 Appendix to Chapter 3

8.3.1 Variational Monte Carlo

8.3.2 Monte Carlo expectation values

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

$$\langle \hat{B} \rangle = \frac{\langle \Psi_\theta | \hat{B} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle} = \sum_{\{\sigma\}} P_\theta(\sigma) B_L(\sigma) \ , \tag{8.33}$$

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:

$$\bar{B} = \frac{1}{M} \sum_{i=1}^{M} B_L(\sigma_i) \ , \tag{8.34}$$

with $\{\sigma_1, \sigma_2, \ldots, \sigma_M\}$ generated from the distribution $P_\theta(\sigma)$ (see Appendix 8.3.3). 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:

$$B_L(\sigma) = \sum_{\{\sigma'\}} \langle \sigma | \hat{B} | \sigma' \rangle \, \frac{\Psi_\theta(\sigma')}{\Psi_\theta(\sigma)} \ . \tag{8.35}$$

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.

8.3.3 Metropolis Algorithm

The Metropolis algorithm allows the generation of a Markov Chain [119] of configurations $\{\sigma_1, \sigma_2, \ldots, \sigma_M\}$ 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* [119].

2. Evaluate the log-acceptance ratio of the proposed move:

$$\log[A(\sigma', \sigma)] = \min\left(0, \log\left[\frac{P_\theta(\sigma')}{P_\theta(\sigma)}\right]\right) \ , \tag{8.36}$$

where

$$\log\left[\frac{P_\theta(\sigma')}{P_\theta(\sigma)}\right] = 2\,\mathrm{Re}\{\mathrm{Log}[\Psi_\theta(\sigma')]\} - 2\,\mathrm{Re}\{\mathrm{Log}[\Psi_\theta(\sigma)]\} \ , \tag{8.37}$$

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 $\mathrm{Log}[\Psi_\theta(\sigma)]$. This is useful from a practical standpoint to avoid numerical issues, such as underflow and overflow, when evaluating the non-normalized wave function.

8.4 Appendix to Chapter 4

8.4.1 Lattice and symmetries of the Shastry-Sutherland model

The Shastry-Sutherland lattice is shown in Fig. 8.1, where each site is labeled by the Cartesian coordinate $\mathbf{r} = a/2(x, y)$, where x and y are integers and $a = 2$ is the length of the primitive lattice vectors $\mathbf{a}_1 = (a, 0)$ and $\mathbf{a}_2 = (0, a)$. The lattice is invariant under translations $T_{\mathbf{a}_1} : (x, y) \to (x + a, y)$ and $T_{\mathbf{a}_2} : (x, y) \to (x, y + a)$. This symmetry can be easily encoded in the Transformer architecture by taking as input patches the four spins in an empty plaquette (i.e., plaquettes with no J' bonds), which constitute the unit cell and then choosing the translationally invariant attention weights, namely $\alpha_{i,j} = \alpha_{i-j}$. In addition, the lattice is invariant under the rotation with respect to the center of the empty plaquette at the origin of the lattice $R_{\pi/2} : (x, y) \to (-y + 1, x)$ and the diagonal reflection $\sigma_{xy} : (x, y) \to (y + 1, x - 1)$. For the ground state, which lies in the $\mathbf{K}a = (0, 0)$ sector, all these symmetries can be enforced by a projector operator, leading to a total-symmetric state [20, 206, 279]:

$$\tilde{\Psi}_\theta(\sigma) = \sum_{\rho, \mathcal{R}} \Psi_\theta(\rho \mathcal{R} \sigma), \tag{8.38}$$

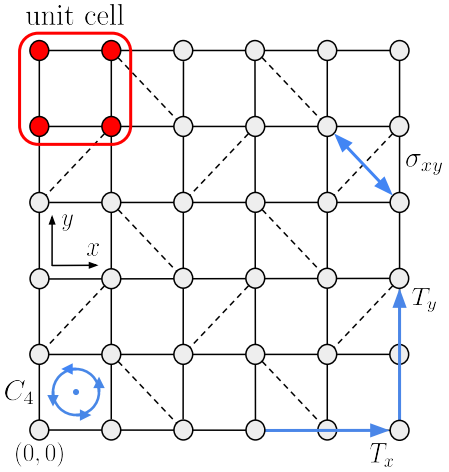


Figure 8.1: The (nearest-neighbor) coupling J is denoted by solid lines and (next-nearest-neighbor one) J' by dashed lines. The standard unit cell contains 4 sites, implying translations T_x and T_y (along x and y axis) by 2 lattice points. The point-group symmetries, C_4 rotations and σ_{xy} reflection, are also shown.

where $\rho \in \{\mathbb{I}, \sigma_{xy}\}$ and $\mathcal{R} \in \{\mathbb{I}, R_{\pi/2}, R^2_{\pi/2}, R^3_{\pi/2}\}$. Notice that the sum is over a fixed number of terms and does *not* scale with the size of the system. In general, this procedure gets an improvement in the accuracy of the variational state, which is difficult to obtain by just increasing the number of variational parameters. The numerical simulations shown in this work are performed with the symmetrized state in Eq. (8.38).

8.4.2 Extrapolations details of the order parameters of the Shastry-Sutherland model

Here, we provide further details on the extrapolation procedures used to obtain the final values of the order parameters presented in Fig. 5.4 of the main text. In the upper panels of Fig. 8.2, we show the order parameters, namely the square magnetization $m^2(L)$ and the plaquette order parameter $m_p(L)$, plotted as a function of the number of layers n_l, extrapolating their values for a network with $n_l \to \infty$. For $L = 6$, these numerical extrapolations show excellent agreement with exact diagonalization results. Furthermore, the extrapolated values exhibit minimal deviation from those obtained with $n_l = 8$ layers, underscoring the consistency of the results. To further assess the robustness of the calculations, in the lower panels of Fig. 8.2 the extrapolations are also performed as a function of the variance of the energy. Notably, this alternative method, which does not depend on the specific structure of the variational state, yields results consistent with the layer-based

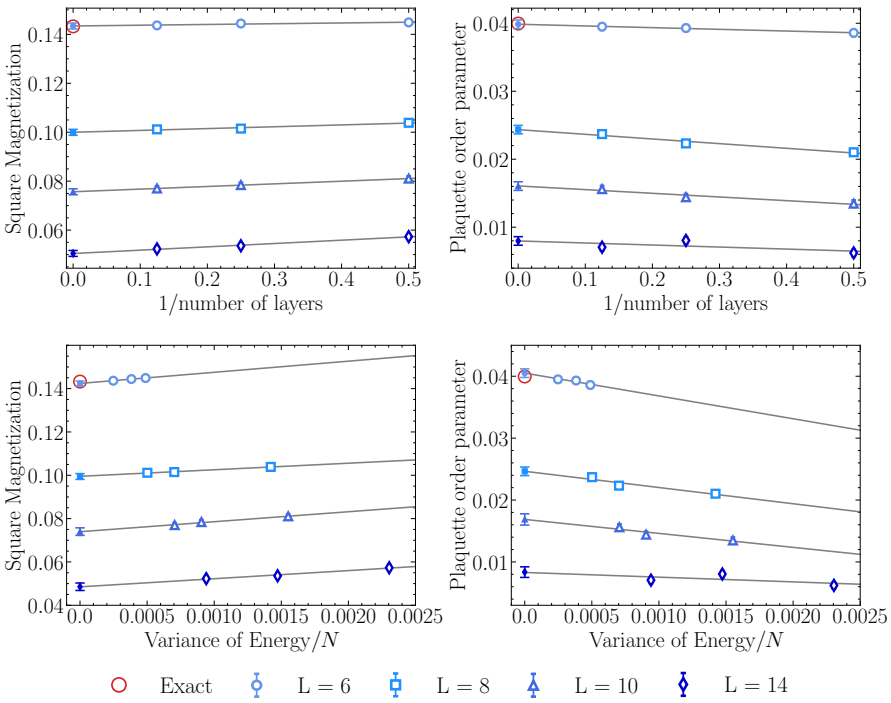


Figure 8.2: Upper panels: the values of the square magnetization $m^2(L)$ (left panel) and of the plaquette order parameter $m_p(L)$ (right panel) as a function of the number of layers $n_l = 2, 4, 8$ and for different lattice sizes from $L = 6$ to $L = 14$ at $J/J' = 0.8$. Lower panels: the same quantities plotted as function of the variance of the energy (divided by N). Extrapolated results for the limit of an infinite number of layers and in the limit of zero variance are displayed as filled symbols, with exact values for $L = 6$ shown for comparison (red circles). Error bars on the extrapolated values are determined using resampling techniques under the assumption of Gaussian noise.

extrapolations. Finally, in Fig. 8.3, we cross-validate the results by utilizing both a second-order polynomial in $1/L$ [see Eq. (5.5)] and the critical form described in Eq. (5.6) for the extrapolations of the square magnetization in the thermodynamic limit (as a function of $1/L$). We repeat the extrapolations considering each value of the number of layers $n_l = 2$, 4, and 8. The resulting fitting curves exhibit remarkably similar behaviors, further confirming the consistency and reliability of the extrapolated results.

Table 8.1 reports the ground-state variational energies per site (in units of J') for system sizes ranging from $L = 6$ to $L = 14$, computed using a Transformer architecture with n_l layers, where $n_l = 2$, 4, and 8. The last column shows the extrapolated energies for an infinite number of layers, obtained through variance

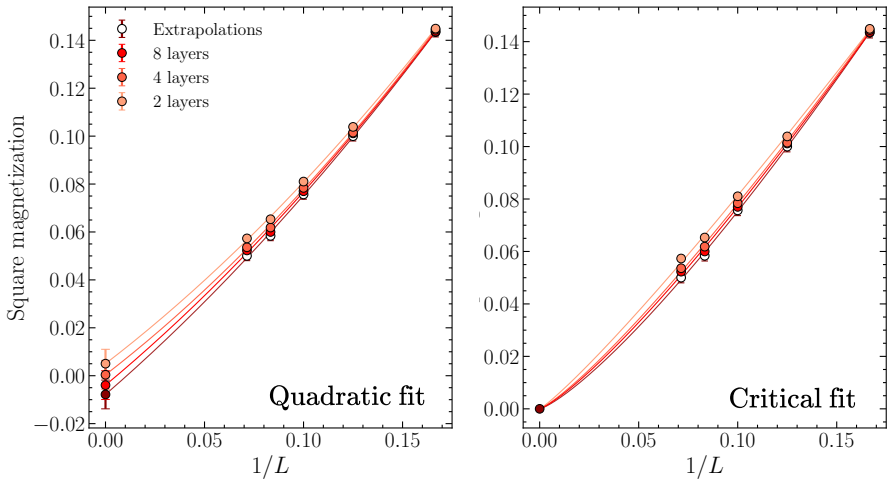


Figure 8.3: Size scaling of the square magnetization $m^2(L)$ as a function of $1/L$ from $L = 6$ up to $L = 14$ at $J'/J = 0.8$. The numerical data encompasses varying numbers of layers, specifically from $n_l = 2$ to $n_l = 8$, along with the extrapolated values for an infinite number of layers (see Fig. 8.2). The curves for the extrapolations in the thermodynamic limit are performed using as fitting curve a second-order polynomial in $1/L$ [see Eq. (5.5)] (left panel) and the critical form of Eq. (5.6) (right panel). The error bars of the extrapolated values in the thermodynamic limit are obtained with resampling techniques with gaussian noise.

extrapolation technique [41]. Additionally, the ground-state energies for $n_l = 8$ are provided for the largest system sizes considered in this work, $L = 16$ and $L = 18$.

8.4.3 ViT study of the Shastry-Sutherland model on large lattice sizes

The ViT architecture outlined in Section 4.1, due to the fully connected structure of the attention mechanism [see Eq. (3.5)], exhibits a computational complexity that scales quadratically with the input sequence length. This quadratic scaling leads to substantial computational costs when increasing the size of the system. Here, we propose two alternative approaches to mitigate this computational challenge.

The first approach utilizes *local* attention mechanisms, which compute the updated representation of a patch based only on its neighboring patches. This method is employed in our simulations of the 16×16 and 18×18 lattice systems. Specifically, for these larger sizes, we utilize a network with hyperparameters $h = 12$, $d = 72$, $n_l = 8$, and a local attention window size of 5×5 patches. In this case, we do not perform the extrapolations of the order parameters in the number of layers: the robustness of the values obtained for a number of layers $n_l = 8$ is justified

L	2 layers	4 layers	8 layers	Extrap.
6	-0.4516642	-0.4516991	-0.4517072	-0.451750
8	-0.449641	-0.449802	-0.449829	-0.44995
10	-0.449062	-0.449221	-0.449329	-0.44947
12	-0.448861	-0.449013	-0.449078	-0.44931
14	-0.448551	-0.448812	-0.448929	-0.44920
16			-0.448881	
18			-0.448859	

Table 8.1: Ground-state variational energy (in unit of J') for different number of layers n_l at $J/J' = 0.8$. The extrapolated values obtained by variance extrapolation [206] for an infinite number of layers are also reported. The Monte Carlo error due to finite sampling effects is on the last digit. In the case of a 6×6 lattice, the ground-state energy per site from exact diagonalization is $E = -0.4517531$.

by the analysis presented in Appendix 8.4.2. We emphasize that the use of local attention preserves the translational invariance of the variational state.

The second method to reduce the computational complexity with the input sequence length is to increase the patch size b. This method has the advantage of preserving global connections of the patches in each layer but has the drawback of breaking translational symmetries, which must subsequently be restored, as with other symmetries (see Appendix 8.4.1).

In Fig. 8.4, we present the optimization curves of the variational energy for a 16×16 lattice, employing a neural network architecture with hyperparameters $h = 12$, $d = 72$, and $n_l = 12$, and utilizing a patch size of $b = 4$. The optimization proceeds in stages: an initial phase of 10^4 steps is performed without enforcing symmetries, during which the ViT state retains only translational symmetry among patches. This first stage is followed by the restoration of translational symmetry over 4×10^3 steps. Subsequently, C_4 rotational symmetry and reflection symmetry are imposed, optimized over 2×10^3 and 10^3 steps, respectively. For comparison, horizontal dashed lines represent the final energies obtained using a local attention approach with a patch size of $b = 2$ and hyperparameters $h = 12$, $d = 72$, $n_l = 8$. The close agreement between the final energies across both approaches highlights the robustness and consistency of the variational results across the two different setups.

8.4.4 Hidden Representations in Frustrated scenario

In subsection 4.1.3, we discussed the role of the hidden representations in determining the effectiveness of our variational *Ansatz* [see Eq. (4.1)]. In Fig. 8.5, we

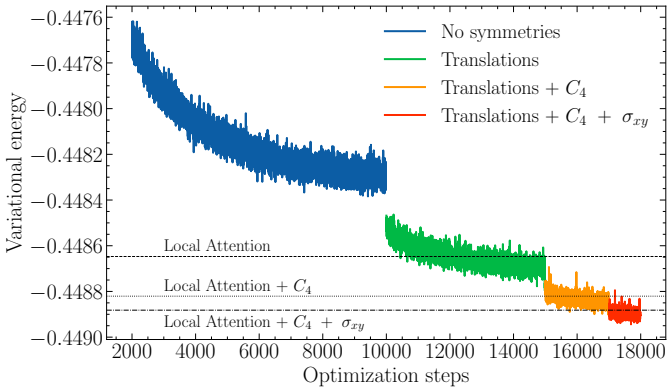


Figure 8.4: Variational energy as a function of optimization steps for a ViT optimized to approximate the ground state of a 16×16 lattice at coupling ratio $J/J' = 0.8$. The hyperparameters of the network are set as follows: attention heads $h = 12$, embedding dimension $d = 72$, number of layers $n_l = 12$, and a patch size of $b = 4$. For enhanced visual clarity, the initial 2×10^3 steps have been omitted from the plot. Horizontal dashed lines indicate the final energies obtained using a local attention mechanism with a patch size of $b = 2$ and a window size restricted to 5×5 patches. For this local attention setup, the network's hyperparameters are $h = 12$, $d = 72$, and $n_l = 8$. In both setups, the optimization protocol employs a sample size of $M = 6 \times 10^3$, a learning rate of $\tau = 0.03$, and a diagonal shift regularization parameter of $\lambda = 10^{-4}$.

compare UMAP projections for the unfrustrated case and the frustrated case with $J/J' = 0.7$. In these plots, the hidden representations $\boldsymbol{z}_i$ are visualized with colors assigned according to the predicted logarithmic amplitude of the associated configurations $\boldsymbol{\sigma}_i$, which is normalized relative to the maximum amplitude within the sample. Importantly, we observe that even for a generic frustration ratio J/J' , a clustering structure consistently emerges in the feature space at the end of the energy minimization. This outcome validates the representation learning framework that inspired the design of our *Ansatz*. Despite starting from the same random initialization of the variational parameters, resulting in all configurations being concentrated within a single cluster (see Fig. 4.4a), the feature spaces obtained after the energy minimization exhibit significant differences. For the unfrustrated case, there is a global gradient in the amplitudes. Points in the bottom clusters correspond to configurations with a small number of spin flips relative to the Néel state, resulting in larger amplitudes (see Fig. 4.4). Moving toward the top clusters, the number of spin flips increases, leading to a decay in amplitude. In contrast, for the frustrated case, the arrangement of configurations in the feature space is significantly different. Configurations mapped within the same cluster display

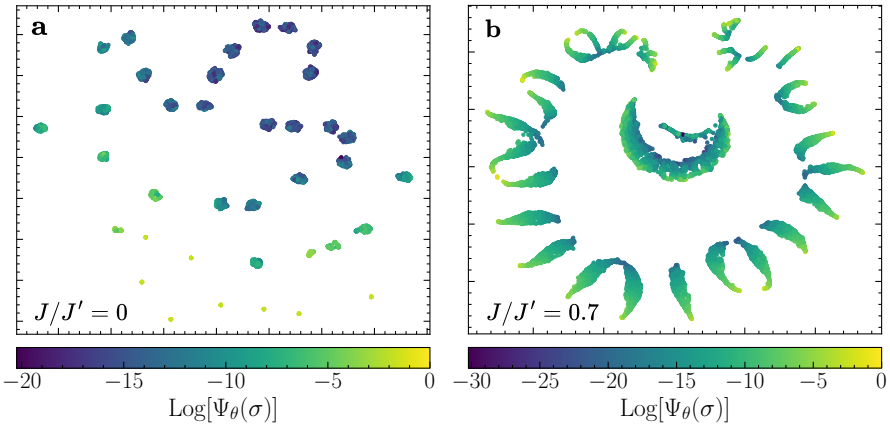


Figure 8.5: UMAP projections of the hidden representations, at the end of the variational energy minimization, for the unfrustrated [panel (a)] and frustrated case with $J/J' = 0.7$ [panel (b)] on a 8×8 cluster. Points are colored according to the log-amplitude of the corresponding spin configurations, normalized with respect to the maximum amplitude within the sample.

very different amplitudes, with each cluster exhibiting its own internal gradient of amplitudes. This complication makes interpreting the clustering structure more challenging compared to the unfrustrated scenario.

8.5 Appendix to Chapter 7

8.5.1 FNQS expectation values

Given a set of operators $\hat{A}_{\boldsymbol{\gamma}}$ parametrized by the couplings $\boldsymbol{\gamma}$, its ensemble average over the distribution $\mathcal{P}(\boldsymbol{\gamma})$ is expressed as:

$$A(\theta) = \int d\boldsymbol{\gamma} \mathcal{P}(\boldsymbol{\gamma}) \frac{\langle \psi_\theta(\boldsymbol{\gamma}) | \hat{A}_{\boldsymbol{\gamma}} | \psi_\theta(\boldsymbol{\gamma}) \rangle}{\langle \psi_\theta(\boldsymbol{\gamma}) | \psi_\theta(\boldsymbol{\gamma}) \rangle} \ . \tag{8.39}$$

This expectation value can be stochastically evaluated using a set of $\mathcal{R}$ couplings $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_\mathcal{R}\}$ sampled from the probability distribution $\mathcal{P}(\boldsymbol{\gamma})$ as:

$$A(\theta) \approx \frac{1}{\mathcal{R}} \sum_{k=1}^{\mathcal{R}} \frac{\langle \psi_\theta(\boldsymbol{\gamma}_k) | \hat{A}_{\boldsymbol{\gamma}_k} | \psi_\theta(\boldsymbol{\gamma}_k) \rangle}{\langle \psi_\theta(\boldsymbol{\gamma}_k) | \psi_\theta(\boldsymbol{\gamma}_k) \rangle} \ . \tag{8.40}$$

Each term in the sum of Eq. (8.40) can be rewritten as:

$$\frac{\langle \psi_\theta(\boldsymbol{\gamma}_k) | \hat{A}_{\boldsymbol{\gamma}_k} | \psi_\theta(\boldsymbol{\gamma}_k) \rangle}{\langle \psi_\theta(\boldsymbol{\gamma}_k) | \psi_\theta(\boldsymbol{\gamma}_k) \rangle} = \sum_{\boldsymbol{\sigma}} p_\theta(\boldsymbol{\sigma} | \boldsymbol{\gamma}_k) \frac{\langle \boldsymbol{\sigma} | \hat{A}_{\boldsymbol{\gamma}_k} | \psi_\theta(\boldsymbol{\gamma}_k) \rangle}{\langle \boldsymbol{\sigma} | \psi_\theta(\boldsymbol{\gamma}_k) \rangle} \ . \tag{8.41}$$

where we have defined the probability distribution $p_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma}_k) = |\psi_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma}_k)|^2 / \langle\psi_\theta(\boldsymbol{\gamma}_k)|\psi_\theta(\boldsymbol{\gamma}_k)\rangle$. In the Variational Monte Carlo (VMC) framework [119], this expectation value can be further estimated stochastically over a set of M_k physical configurations $\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{M_k}\}$ sampled according to the probability distribution $p_\theta(\boldsymbol{\sigma}|\boldsymbol{\gamma}_k)$:

$$\bar{A}_k = \frac{1}{M_k} \sum_{j=1}^{M_k} \frac{\langle \boldsymbol{\sigma}_j | \hat{A}_{\boldsymbol{\gamma}_k} | \psi_\theta(\boldsymbol{\gamma}_k)\rangle}{\langle \boldsymbol{\sigma}_j | \psi_\theta(\boldsymbol{\gamma}_k)\rangle} \ . \tag{8.42}$$

In the calculations performed in this work, we set an equal number of samples for each system, $M_k = M/\mathcal{R}$, independent of k, where M is the total number of samples in the extended space of all systems. See to Ref. [119] for further details on the VMC framework.

8.5.2 Derivation of the Stochastic Reconfiguration equation for multiple systems

To derive the Stochastic Reconfiguration (SR) equations for multiple systems, we employ the paradigm of the Time-Dependent Variational Principle (TDVP) [280]. Consider a system characterized by coupling parameters $\boldsymbol{\gamma}$. We begin by defining the fidelity between the state evolved in imaginary time under the exact Hamiltonian for a time-step ε, namely $e^{-\varepsilon \hat{H}_{\boldsymbol{\gamma}}} |\psi_{\theta(\tau)}(\boldsymbol{\gamma})\rangle$, and the corresponding variationally evolved state $|\psi_{\theta(\tau)+\varepsilon\dot{\theta}(\tau)}(\boldsymbol{\gamma})\rangle$. Here, $\dot{\theta}_\alpha(\tau)$ denotes the derivative of the α-th variational parameter with respect to imaginary time τ, with $\alpha = 1, \ldots, P$ and P the total number of parameters. For simplicity in the notation, in the following, we omit the explicit time dependence of the variational parameters.

The fidelity is then given by:

$$f^2(\boldsymbol{\gamma}) = \frac{|\langle\psi_\theta(\boldsymbol{\gamma})|e^{-\varepsilon\hat{H}_{\boldsymbol{\gamma}}}|\psi_{\theta+\varepsilon\dot{\theta}}(\boldsymbol{\gamma})\rangle|^2}{\langle\psi_\theta(\boldsymbol{\gamma})|e^{-\varepsilon\hat{H}_{\boldsymbol{\gamma}}}|\psi_\theta(\boldsymbol{\gamma})\rangle\langle\psi_{\theta+\varepsilon\dot{\theta}}(\boldsymbol{\gamma})|\psi_{\theta+\varepsilon\dot{\theta}}(\boldsymbol{\gamma})\rangle} \ . \tag{8.43}$$

Expanding this expression, assuming real-valued variational parameters and neglecting third-order contributions in ε, we obtain:

$$f^2(\boldsymbol{\gamma}) \approx 1 - \varepsilon^2 \left[\dot{\boldsymbol{\theta}}^T \boldsymbol{G}(\boldsymbol{\gamma}) + \dot{\boldsymbol{\theta}}^T S(\boldsymbol{\gamma}) \dot{\boldsymbol{\theta}} + \mathrm{Var}(\hat{H}_{\boldsymbol{\gamma}})\right] \ . \tag{8.44}$$

Here, $G_\alpha(\boldsymbol{\gamma})$ is the gradient of the energy defined in Eq. (7.4) and $S_{\alpha,\beta}(\boldsymbol{\gamma})$ is the real part of the quantum geometric tensor defined in Eq. (7.5). To extend this to an ensemble of systems, we define the global fidelity as the ensemble average of the fidelity over the distribution $\mathcal{P}(\boldsymbol{\gamma})$ as $\mathcal{F}^2 = \int d\boldsymbol{\gamma}\ \mathcal{P}(\boldsymbol{\gamma}) f^2(\boldsymbol{\gamma})$. Substituting Eq. (8.44) into the previous expression yields

$$\mathcal{F}^2 \approx 1 - \varepsilon^2 \left[\dot{\boldsymbol{\theta}}^T \mathcal{G}(\boldsymbol{\gamma}) + \dot{\boldsymbol{\theta}}^T \mathcal{S}(\boldsymbol{\gamma}) \dot{\boldsymbol{\theta}} + \int d\boldsymbol{\gamma} \mathcal{P}(\boldsymbol{\gamma}) \mathrm{Var}(\hat{H}_{\boldsymbol{\gamma}})\right] \ , \tag{8.45}$$

where the ensemble-averaged quantities are defined as $\mathcal{G}_\alpha = \int d\boldsymbol{\gamma}\mathcal{P}(\boldsymbol{\gamma})G_\alpha(\boldsymbol{\gamma})$ and $\mathcal{S}_{\alpha,\beta} = \int d\boldsymbol{\gamma}\mathcal{P}(\boldsymbol{\gamma})S_{\alpha,\beta}(\boldsymbol{\gamma})$ [see Section 7.2.1] .

The TDVP equations for the ensemble are obtained by maximizing the global fidelity in Eq. (8.45) with respect to $\dot{\boldsymbol{\theta}}$, leading to the linear system $\mathcal{S}\dot{\boldsymbol{\theta}} = -\frac{1}{2}\boldsymbol{\mathcal{G}}$. This differential equation can then be integrated numerically. In ground-state applications, it is common to employ the simple Euler scheme, which approximates the time derivative as $\dot{\theta}_\alpha(\tau) \approx [\theta(\tau+\eta) - \theta(\tau)]/\eta$. Here, τ denotes the imaginary time parameterizing the dynamics, and η is the integration time step used to discretize the evolution. Based on these results, the SR updates are conventionally defined by removing the factor of 1/2 [281, 282] leading to $\boldsymbol{\delta\theta} = -\eta\mathcal{S}^{-1}\boldsymbol{\mathcal{G}}$, where we have defined $\delta\theta_\alpha = \theta_\alpha(\tau+\eta) - \theta_\alpha(\tau)$. It is important to consider that the matrix $\mathcal{S}$ may possess extremely small or even negligible eigenvalues. As a result, directly computing its inverse can lead to numerical instabilities [119]. To mitigate these potential issues, we adopt a regularized update scheme of the form:

$$\boldsymbol{\delta\theta} = -\eta\left(\mathcal{S} + \lambda\mathbb{I}_P\right)^{-1}\boldsymbol{\mathcal{G}} \ , \tag{8.46}$$

where η acts as a learning rate controlling the update magnitude during the optimization process, and $\lambda > 0$ is a regularization parameter introduced to ensure the invertibility and numerical stability of the matrix $\mathcal{S}$ matrix [37, 119]. The resulting updates are also reported in Eq. (7.3) of Section 7.2.1.

8.5.3 FNQS systematic energy improvement

In this section, we systematically increase the expressivity of the FNQS by varying the number of parameters and assess the resulting performance using the V-score [283], a metric known to correlate with the accuracy of the wave function. The V-score for the system $\hat{H}_{\boldsymbol{\gamma}}$ is defined as

$$\text{V-score}(\boldsymbol{\gamma}) = N\frac{\langle H^2_{\boldsymbol{\gamma}}\rangle_{\boldsymbol{\gamma}} - \langle H_{\boldsymbol{\gamma}}\rangle^2_{\boldsymbol{\gamma}}}{\langle H_{\boldsymbol{\gamma}}\rangle^2_{\boldsymbol{\gamma}}} \ . \tag{8.47}$$

A key advantage of the V-score is that it can be computed without requiring access to exact reference energies, making it particularly suitable for benchmarking variational methods on large systems [283].

The goal of this analysis is to demonstrate that the accuracy of the FNQS can be systematically improved by increasing the size of the neural network, which we control by adjusting the number of layers. We carry out this study on the J_1–J_2 Heisenberg model on the square lattice, a widely established benchmark, obtained by setting $J_3 = 0$ in Eq. (7.10). Specifically, we consider system sizes ranging from $L = 6$ to $L = 12$, and vary the number of layers from $n_l = 2$ to $n_l = 8$. The FNQS

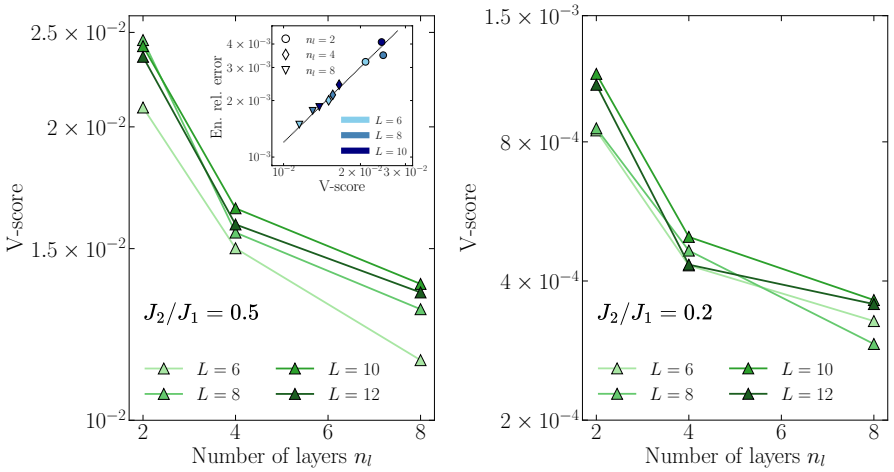


Figure 8.6: V-score for the J_1–J_2 Heisenberg model on the square lattice as a function of the number of layers n_l of the FNQS, for increasing the system size from $L = 6$ to $L = 12$. The left panel shows results for the highly frustrated point $J_2/J_1 = 0.5$, while the right panel corresponds to the less frustrated point $J_2/J_1 = 0.2$. The inset in the left panel displays the correlation between the V-score and the relative energy error at $J_2/J_1 = 0.5$ for system sizes $L = 6$ to $L = 10$. Reference energies for $L = 8$ and $L = 10$ are obtained via zero-variance extrapolation from Refs. [41, 219].

is optimized across $\mathcal{R} = 1000$ coupling values, uniformly sampled in the interval $J_2/J_1 \in [0, 1]$. Importantly, for a fixed network architecture and total number of Monte Carlo samples M, the computational cost of each simulation remains independent of $\mathcal{R}$. Additionally, while the computational complexity of the ViT scales quadratically with the input sequence length, this cost can be mitigated through various strategies [21].

After optimizing the FNQS, we focus on two representative values of the frustration ratio: $J_2/J_1 = 0.5$ and $J_2/J_1 = 0.2$. In the left panel of Fig. 8.6, we present results for $J_2/J_1 = 0.5$, a highly frustrated point in the phase diagram [20]. At this value, high-precision ground-state energy estimates obtained via zero-variance extrapolation are available for system sizes beyond $L = 6$ [41, 219], enabling us to validate the correlation between the relative energy error and the V-score, as shown in the inset of the left panel of Fig. 8.6. In the right panel of Fig. 8.6, we report the V-score for $J_2/J_1 = 0.2$, a less frustrated point closer to the unfrustrated Heisenberg model. Due to the reduced complexity of the ground state in this region, the V-scores are systematically lower compared to the highly frustrated case

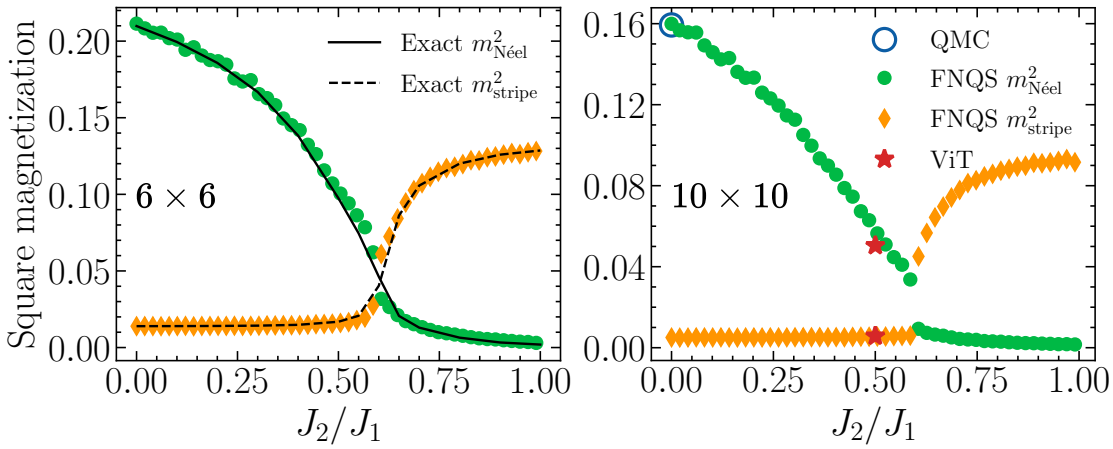


Figure 8.7: Square magnetization corresponding to the Néel (m^2_{Nel}) and stripe (m^2_{stripe}) order as a function of the frustration ratio J_2/J_1. The left panel corresponds to a 6×6 lattice, in which the FNQS results at $J_3/J_1 = 0$ are compared with exact diagonalization calculations (solid and dashed black lines). The right panel corresponds to a 10×10 lattice, in which the variational results are compared with Quantum Monte Carlo (QMC, blue circles) at $J_2/J_1 = 0$, and with Vision Transformer architecture (ViT, red stars) at $J_2/J_1 = 0.5$.

at $J_2/J_1 = 0.5$. By analyzing system sizes up to $N = 144$ spins for both frustration ratios, we find that the V-score, and thus the accuracy, systematically improves when increasing the number of parameters in the network, while marginally deteriorating with the system size for a fixed architecture. These results highlight the size-consistency of the ViT parametrization of the FNQS, even when the model is simultaneously optimized across multiple systems. We remark that further improvements in accuracy are achievable by explicitly incorporating the symmetries of the underlying Hamiltonian, although this lies beyond the scope of the present analysis.

8.5.4 FNQS study of the structure factor for $J_3/J_1 = 0$

In this section, we evaluate the accuracy of the FNQS trained within the parameter region $J_2/J_1 \in [0, 1.0]$ and $J_3/J_1 \in [0, 0.6]$ of the J_1-J_2-J_3 Heisenberg model (see Section 7.3.2), focusing on the line $J_3/J_1 = 0$.

The left panel of Fig. 8.7 shows the results for a 6×6 lattice, where we compute the Néel and stripe order parameters, respectively $m^2_{\text{Néel}}$ and m^2_{stripe}, as defined in Section 7.3.2. Our FNQS predictions are in excellent agreement with exact

diagonalization results, demonstrating the reliability of the FNQS architecture.

In the right panel of Fig. 8.7, we extend this analysis to a 10×10 lattice, using the same setup as in Section 7.3.2. Since exact diagonalization is infeasible at this system size, we benchmark FNQS predictions against Quantum Monte Carlo (QMC) data at the unfrustrated point $J_2/J_1 = 0.0$ [190] and against results from a state-of-the-art ViT architecture trained from scratch at $J_2/J_1 = 0.5$ [37].

8.5.5 Derivation of the generalized fidelity susceptibility

The fidelity $F^2(\boldsymbol{\gamma}, \boldsymbol{\varepsilon})$ defined in Eq. (7.7) can be expanded in a Taylor series around $\boldsymbol{\varepsilon} = 0$ as

$$F^2(\boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = 1 - \sum_{i,j=1}^{N_c} \varepsilon_i \varepsilon_j \chi_{ij}(\boldsymbol{\gamma}) + O(|\boldsymbol{\varepsilon}|^3), \tag{8.48}$$

where N_c denotes the number of couplings.

The leading non-zero contribution to fidelity is given by the real part of the *quantum geometric tensor* with respect to couplings $\boldsymbol{\gamma}$ [226, 227], as defined in Eq. (7.8).

Taking the logarithm of Eq. (8.48), we obtain

$$\ln[F(\boldsymbol{\gamma}, \boldsymbol{\varepsilon})] = -\frac{1}{2} \sum_{i,j=1}^{N_c} \varepsilon_i \varepsilon_j \chi_{ij}(\boldsymbol{\gamma}) + O(|\boldsymbol{\varepsilon}|^3). \tag{8.49}$$

Differentiating twice with respect to $\boldsymbol{\varepsilon}$, we find the Hessian matrix

$$\frac{\partial^2 \ln F(\boldsymbol{\gamma}, \boldsymbol{\varepsilon})}{\partial \varepsilon_i \partial \varepsilon_j} = -\chi_{ij}(\boldsymbol{\gamma}) + O(|\boldsymbol{\varepsilon}|). \tag{8.50}$$

At the end, evaluating Eq. (8.50) at $\boldsymbol{\varepsilon} = 0$ yields the *generalized fidelity susceptibility*, as defined in Eq. (7.6).

Bibliography

1. Vaswani, A. *et al. Attention Is All You Need* Dec. 2017. http://arxiv.org/abs/1706.03762.
2. Tay, Y., Dehghani, M., Bahri, D. & Metzler, D. Efficient Transformers: A Survey. https://arxiv.org/abs/2009.06732 (Sept. 2020).
3. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* 2019. arXiv: 1810.04805.
4. Howard, J. & Ruder, S. *Universal Language Model Fine-tuning for Text Classification* 2018. arXiv: 1801.06146 [cs.CL].
5. Radford, A., Narasimhan, K., *et al. Improving language understanding by generative pre-training* 2018.
6. Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33,** 1877–1901 (2020).
7. OpenAI. *GPT-4 Technical Report* 2024. arXiv: 2303.08774 [cs.CL].
8. Hyvärinen, A. Consistency of Pseudolikelihood Estimation of Fully Visible Boltzmann Machines. *Neural Computation* **18,** 2283–2292. ISSN: 0899-7667. https://doi.org/10.1162/neco.2006.18.10.2283 (Oct. 2006).
9. Ravikumar, P., Wainwright, M. J. & Lafferty, J. D. High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics* **38,** 1287–1319. https://doi.org/10.1214/09-AOS691 (2010).
10. Aurell, E. & Ekeberg, M. Inverse Ising Inference Using All the Data. *Phys. Rev. Lett.* **108,** 090201. https://link.aps.org/doi/10.1103/PhysRevLett.108.090201 (9 Mar. 2012).
11. Saad, D. & Solla, S. Exact Solution for On-Line Learning in Multilayer Neural Networks. *Phys. Rev. Lett.* **74,** 4337–4340 (1995).
12. Farnia, F., Zhang, J. & Tse, D. *A Spectral Approach to Generalization and Optimization in Neural Networks* in *ICLR* (2018).

13. Pérez, G. V., Camargo, C. Q. & Louis, A. A. *Deep learning generalizes because the parameter-function map is biased towards simple functions* in *International Conference on Learning Representations* (OpenReview.net, 2019).

14. Kalimeris, D. *et al. SGD on Neural Networks Learns Functions of Increasing Complexity* in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (eds Wallach, H. M. *et al.*) (2019), 3491–3501.

15. Rahaman, N. *et al. On the Spectral Bias of Neural Networks* in *Proc. of ICML* (eds Chaudhuri, K. & Salakhutdinov, R.) **97** (PMLR, 2019), 5301–5310.

16. Becca, F. & Sorella, S. *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University Press, 2017).

17. Carleo, G. & Troyer, M. Solving the Quantum Many-Body Problem with Artificial Neural Networks. *Science* **355,** 602–606. http://arxiv.org/abs/1606.02318 (Feb. 2017).

18. Glasser, I., Pancotti, N., August, M., Rodriguez, I. & Cirac, J. Neural-Network Quantum States, String-Bond States, and Chiral Topological States. *Phys. Rev. X* **8,** 011006. https://link.aps.org/doi/10.1103/PhysRevX.8.011006 (1 Jan. 2018).

19. Lange, H., Van de Walle, A., Abedinnia, A. & Bohrdt, A. From architectures to applications: a review of neural quantum states. *Quantum Science and Technology* **9,** 040501. https://dx.doi.org/10.1088/2058-9565/ad7168 (Sept. 2024).

20. Nomura, Y. & Imada, M. 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. https://link.aps.org/doi/10.1103/PhysRevX.11.031034 (3 Aug. 2021).

21. Viteritti, L. L., Rende, R., Parola, A., Goldt, S. & Becca, F. Transformer wave function for two dimensional frustrated magnets: Emergence of a spin-liquid phase in the Shastry-Sutherland model. *Phys. Rev. B* **111,** 134411. https://link.aps.org/doi/10.1103/PhysRevB.111.134411 (13 Apr. 2025).

22. Chen, A. & Heyl, M. *Efficient optimization of deep neural quantum states toward machine precision* 2023. arXiv: 2302.01941 [cond-mat.dis-nn].

23. Roth, C., Szabó, A. & MacDonald, A. H. High-accuracy variational Monte Carlo for frustrated magnets with deep neural networks. *Phys. Rev. B* **108,** 054410. https://link.aps.org/doi/10.1103/PhysRevB.108.054410 (5 Aug. 2023).

24. Denis, Z. & Carleo, G. Accurate neural quantum states for interacting lattice bosons. *arXiv preprint arXiv:2404.07869.* arXiv: 2404.07869 [quant-ph] (2024).

25. Robledo Moreno, J., Carleo, G., Georges, A. & Stokes, J. Fermionic wave functions from neural-network constrained hidden states. *Proceedings of the National Academy of Sciences* **119.** ISSN: 1091-6490. http://dx.doi.org/10.1073/pnas.2122059119 (Aug. 2022).

26. Kim, J. *et al.* Neural-network quantum states for ultra-cold Fermi gases. *arXiv preprint arXiv:2305.08831.* arXiv: 2305.08831 [cond-mat.quant-gas] (2023).

27. Pfau, D., Spencer, J. S., Matthews, A. G. D. G. & Foulkes, W. M. C. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2,** 033429. https://link.aps.org/doi/10.1103/PhysRevResearch.2.033429 (3 Sept. 2020).

28. Nys, J., Pescia, G. & Carleo, G. Ab-initio variational wave functions for the time-dependent many-electron Schrödinger equation. *arXiv preprint arXiv:2403.07447.* arXiv: 2403.07447 [cond-mat.str-el] (2024).

29. Choo, K., Neupert, T. & Carleo, G. Two-dimensional frustrated J_1-J_2 model studied with neural network quantum states. *Phys. Rev. B* **100,** 125124. https://link.aps.org/doi/10.1103/PhysRevB.100.125124 (12 Sept. 2019).

30. Liang, X. *et al.* Solving frustrated quantum many-particle models with convolutional neural networks. *Phys. Rev. B* **98,** 104426. https://link.aps.org/doi/10.1103/PhysRevB.98.104426 (10 Sept. 2018).

31. Chen, A., Choo, K., Astrakhantsev, N. & Neupert, T. Neural network evolution strategy for solving quantum sign structures. *Phys. Rev. Research* **4,** L022026. https://link.aps.org/doi/10.1103/PhysRevResearch.4.L022026 (2 May 2022).

32. Roth, C. & MacDonald, A. *Group Convolutional Neural Networks Improve Quantum State Accuracy* May 2021. http://arxiv.org/abs/2104.05085.

33. Hibat-Allah, M., Ganahl, M., Hayward, L., Melko, R. & Carrasquilla, J. Recurrent neural network wave functions. *Phys. Rev. Research* **2,** 023358. https://link.aps.org/doi/10.1103/PhysRevResearch.2.023358 (2 June 2020).

34. Hibat-Allah, M., Melko, R. & Carrasquilla, J. *Supplementing Recurrent Neural Network Wave Functions with Symmetry and Annealing to Improve Accuracy* July 2022. http://arxiv.org/abs/2207.14314.

35. Luo, D. *et al. Gauge Invariant Autoregressive Neural Networks for Quantum Lattice Models* Dec. 2021. http://arxiv.org/abs/2101.07243.

36. Sharir, O., Levine, Y., Wies, N., Carleo, G. & Shashua, A. Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems. *Phys. Rev. Lett.* **124,** 020503. https://link.aps.org/doi/10.1103/PhysRevLett.124.020503 (2 Jan. 2020).

37. Rende, R., Viteritti, L. L., Bardone, L., Becca, F. & Goldt, S. A simple linear algebra identity to optimize large-scale neural network quantum states. *Communications Physics* **7.** ISSN: 2399-3650. http://dx.doi.org/10.1038/s42005-024-01732-4 (Aug. 2024).

38. Shastry, B. & Sutherland, B. Exact ground state of a quantum mechanical antiferromagnet. *Physica B+C* **108,** 1069–1070. https://www.sciencedirect.com/science/article/pii/037843638190838X (1981).

39. Kageyama, H. *et al.* Exact Dimer Ground State and Quantized Magnetization Plateaus in the Two-Dimensional Spin System $SrCu_2(BO_3)_2$. *Phys. Rev. Lett.* **82,** 3168–3171. https://link.aps.org/doi/10.1103/PhysRevLett.82.3168 (15 Apr. 1999).

40. Miyahara, S. & Ueda, K. Exact Dimer Ground State of the Two Dimensional Heisenberg Spin System $SrCu_2(BO_3)_2$. *Phys. Rev. Lett.* **82,** 3701–3704. https://link.aps.org/doi/10.1103/PhysRevLett.82.3701 (18 May 1999).

41. Hu, W.-J., Becca, F., Parola, A. & Sorella, S. Direct evidence for a gapless Z_2 spin liquid by frustrating Néel antiferromagnetism. *Phys. Rev. B* **88,** 060402. https://link.aps.org/doi/10.1103/PhysRevB.88.060402 (6 Aug. 2013).

42. Rende, R. *et al.* Foundation neural-networks quantum states as a unified Ansatz for multiple hamiltonians. *Nature Communications* **16.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-025-62098-x (Aug. 2025).

43. Wang, L., Liu, Y.-H., Imri ška, J., Ma, P. N. & Troyer, M. Fidelity Susceptibility Made Simple: A Unified Quantum Monte Carlo Approach. *Phys. Rev. X* **5,** 031007. https://link.aps.org/doi/10.1103/PhysRevX.5.031007 (3 July 2015).

44. Besag, J. Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society. Series D (The Statistician)* **24,** 179–195. ISSN: 00390526, 14679884. http://www.jstor.org/stable/2987782 (2023) (1975).

45. Cocco, S. & Monasson, R. Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data. *Phys. Rev. Lett.* **106,** 090601. https://link.aps.org/doi/10.1103/PhysRevLett.106.090601 (9 Mar. 2011).

46. Ricci-Tersenghi, F. The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment* **2012,** P08015. https://dx.doi.org/10.1088/1742-5468/2012/08/P08015 (Aug. 2012).

47. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87,** 012707. https://link.aps.org/doi/10.1103/PhysRevE.87.012707 (1 Jan. 2013).

48. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics* **81,** 032601. https://dx.doi.org/10.1088/1361-6633/aa9965 (Jan. 2018).

49. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

50. Dosovitskiy, A. *et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* 2021. arXiv: 2010.11929 [cs.CV].

51. Jumper, J., Evans, R., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 1–11 (Aug. 2021).

52. Bahdanau, D., Cho, K. & Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate* 2016. arXiv: 1409.0473 [cs.CL].

53. Kim, Y., Denton, C., Hoang, L. & Rush, A. M. *Structured Attention Networks* 2017. arXiv: 1702.00887 [cs.CL].

54. Gardner, E. & Derrida, B. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General* **22,** 1983. https://dx.doi.org/10.1088/0305-4470/22/12/004 (June 1989).

55. Seung, H. S., Sompolinsky, H. & Tishby, N. Statistical mechanics of learning from examples. *Phys. Rev. A* **45,** 6056–6091. https://link.aps.org/doi/10.1103/PhysRevA.45.6056 (8 Apr. 1992).

56. Engel, A. & Van den Broeck, C. *Statistical Mechanics of Learning* (Cambridge University Press, 2001).

57. Carleo, G. *et al.* Machine learning and the physical sciences. *Rev. Mod. Phys.* **91,** 045002. https://link.aps.org/doi/10.1103/RevModPhys.91.045002 (4 Dec. 2019).

58. Chung, S., Lee, D. D. & Sompolinsky, H. Classification and Geometry of General Perceptual Manifolds. *Phys. Rev. X* **8,** 031003. https://link.aps.org/doi/10.1103/PhysRevX.8.031003 (3 July 2018).

59. Goldt, S., Mézard, M., Krzakala, F. & Zdeborová, L. Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model. *Phys. Rev. X* **10,** 041044. https://link.aps.org/doi/10.1103/PhysRevX.10.041044 (4 Dec. 2020).

60. Spigler, S., Geiger, M. & Wyart, M. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment* **2020,** 124001 (Dec. 2020).

61. Gerace, F., Loureiro, B., Krzakala, F., Mézard, M. & Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. *Journal of Statistical Mechanics: Theory and Experiment* **2021,** 124013. https://dx.doi.org/10.1088/1742-5468/ac3ae6 (Dec. 2021).

62. Favero, A., Cagnetta, F. & Wyart, M. Locality defeats the curse of dimensionality in convolutional teacher–student scenarios. *Journal of Statistical Mechanics: Theory and Experiment* **2022,** 114012. https://dx.doi.org/10.1088/1742-5468/ac98ab (Nov. 2022).

63. Loureiro, B. *et al.* Learning curves of generic features maps for realistic datasets with a teacher-student model. *Journal of Statistical Mechanics: Theory and Experiment* **2022,** 114001. https://dx.doi.org/10.1088/1742-5468/ac9825 (Nov. 2022).

64. Ingrosso, A. & Goldt, S. Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences* **119,** e2201854119. https://www.pnas.org/doi/abs/10.1073/pnas.2201854119 (2022).

65. Refinetti, M. & Goldt, S. *The dynamics of representation learning in shallow, non-linear autoencoders* in *International Conference on Machine Learning* (2022), 18499–18519.

66. Cui, H. & Zdeborová, L. High-dimensional asymptotics of denoising autoencoders. *Advances in Neural Information Processing Systems* **36** (2024).

67. Potts, R. B. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society* **48,** 106–109 (1952).

68. Wu, F. Y. The Potts model. *Rev. Mod. Phys.* **54,** 235–268. https://link.aps.org/doi/10.1103/RevModPhys.54.235 (1 Jan. 1982).

69. Finn, R. D., Bateman, A., *et al.* Pfam: the protein families database. *Nucleic Acids Research* **42,** D222–D230. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkt1223 (Nov. 2013).

70. Bhattacharya, N. *et al.* Single Layers of Attention Suffice to Predict Protein Contacts. *bioRxiv* (2020).

71. Viteritti, L. L., Rende, R. & Becca, F. Transformer Variational Wave Functions for Frustrated Quantum Spin Systems. *Phys. Rev. Lett.* **130,** 236401. https://link.aps.org/doi/10.1103/PhysRevLett.130.236401 (23 June 2023).

72. Goldt, S. *et al. The Gaussian equivalence of generative models for learning with shallow neural networks* in *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference* (Aug. 2022). https://proceedings.mlr.press/v145/goldt22a.html.

73. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108,** E1293–E1301. https://www.pnas.org/doi/abs/10.1073/pnas.1111471108 (2011).

74. Nguyen, H. C., Zecchina, R. & Berg, J. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics* **66,** 197–261. https://doi.org/10.1080/00018732.2017.1341604 (2017).

75. Livan, G., Novaes, M. & Vivo, P. *Introduction to Random Matrices* https://doi.org/10.10072F978-3-319-70885-0 (Springer International Publishing, 2018).

76. Opper, M. & Kinzel, W. Statistical mechanics of generalization. *Models of Neural Networks III: Association, Generalization, and Representation* (1996).

77. Nakkiran, P. *et al.* Deep double descent: where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment* **2021,** 124003. https://dx.doi.org/10.1088/1742-5468/ac3a74 (Dec. 2021).

78. Eldan, R. & Li, Y. *TinyStories: How Small Can Language Models Be and Still Speak Coherent English?* 2023. arXiv: 2305.07759 [cs.CL].

79. Saxe, A. M., McClelland, J. L. & Ganguli, S. *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks* in *Proc. of ICLR* (eds Bengio, Y. & LeCun, Y.) (2014).

80. Saxe, A., McClelland, J. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* **116,** 11537–11546 (2019).

81. Advani, M., Saxe, A. & Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks* **132,** 428–446 (2020).

82. Mei, S., Montanari, A. & Nguyen, P. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences* **115,** E7665–E7671 (2018).

83. Bowman, B. & Montúfar, G. *Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks* in *Proc. of ICLR* (OpenReview.net, 2022).

84. Bowman, B. & Montúfar, G. Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime. *ArXiv preprint* **abs/2206.02927** (2022).

85. Boursier, E., Pillaud-Vivien, L. & Flammarion, N. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems* **35,** 20105–20118 (2022).

86. Refinetti, M., Ingrosso, A. & Goldt, S. *Neural Networks Trained with SGD Learn Distributions of Increasing Complexity* in *Proceedings of the 40th International Conference on Machine Learning* (JMLR.org, Honolulu, Hawaii, USA, 2023).

87. Bardone, L. & Goldt, S. *Sliding Down the Stairs: How Correlated Latent Variables Accelerate Learning with Neural Networks* in *Proceedings of the 41st International Conference on Machine Learning* (eds Salakhutdinov, R. *et al.*) **235** (PMLR, 2024), 3024–3045.

88. Belrose, N., Pope, Q., Quirke, L., Mallen, A. & Fern, X. Neural Networks Learn Statistics of Increasing Complexity. *arXiv preprint arXiv:2402.04362* (2024).

89. Biderman, S. *et al. Pythia: A suite for analyzing large language models across training and scaling* in *International Conference on Machine Learning* (2023), 2397–2430.

90. Rende, R., Gerace, F., Laio, A. & Goldt, S. *A distributional simplicity bias in the learning dynamics of transformers* in *Advances in Neural Information Processing Systems* (eds Globerson, A. *et al.*) **37** (Curran Associates, Inc., 2024), 96207–96228. `https://proceedings.neurips.cc/paper_files/paper/2024/file/ae6c81a39079ddeb88b034b6ef18c7fe-Paper-Conference.pdf`.

91. Rende, R., Gerace, F., Laio, A. & Goldt, S. Mapping of attention mechanisms to a generalized Potts model. *Phys. Rev. Res.* **6,** 023057. `https://link.aps.org/doi/10.1103/PhysRevResearch.6.023057` (2 2024).

92. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* 2019. arXiv: 1810.04805 [cs.CL].

93. Liu, Y. *et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach* 2019. arXiv: 1907.11692 [cs.CL].

94. Merity, S., Xiong, C., Bradbury, J. & Socher, R. *Pointer Sentinel Mixture Models* 2016. arXiv: 1609.07843 [cs.CL]. https://arxiv.org/abs/1609.07843.

95. Goyal, K., Dyer, C. & Berg-Kirkpatrick, T. *Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis–Hastings* 2022. arXiv: 2106.02736 [cs.LG].

96. Bhattamishra, S., Patel, A., Kanade, V. & Blunsom, P. Simplicity bias in transformers and their ability to learn sparse boolean functions. *arXiv preprint arXiv:2211.12316* (2022).

97. Abbe, E., Bengio, S., Boix-Adsera, E., Littwin, E. & Susskind, J. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems* **36** (2024).

98. Vasudeva, B. *et al.* Simplicity Bias of Transformers to Learn Low Sensitivity Functions. *arXiv preprint arXiv:2403.06925* (2024).

99. Liu, Z., Wang, Y., Kasai, J., Hajishirzi, H. & Smith, N. A. *Probing Across Time: What Does RoBERTa Know and When?* in *Findings of the Association for Computational Linguistics: EMNLP 2021* (2021), 820–842.

100. Choshen, L., Hacohen, G., Weinshall, D. & Abend, O. *The Grammar-Learning Trajectories of Neural Language Models* in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2022), 8281–8297.

101. Cagnetta, F., Petrini, L., Tomasini, U. M., Favero, A. & Wyart, M. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X* **14,** 031001 (2024).

102. Cagnetta, F. & Wyart, M. Towards a theory of how the structure of language is acquired by deep neural networks. *arXiv preprint arXiv:2406.00048* (2024).

103. Garnier-Brun, J., Mézard, M., Moscato, E. & Saglietti, L. How transformers learn structured data: insights from hierarchical filtering. *arXiv:2408.15138* (2024).

104. Radford, A. *et al. Language Models are Unsupervised Multitask Learners* https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf. 2018.

105. Dosovitskiy, A. *et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* June 2021. http://arxiv.org/abs/2010.11929.

106. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Theory of Superconductivity. *Phys. Rev.* **108,** 1175–1204. https://link.aps.org/doi/10.1103/PhysRev.108.1175 (5 Dec. 1957).

107. Laughlin, R. B. Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations. *Phys. Rev. Lett.* **50,** 1395–1398. https://link.aps.org/doi/10.1103/PhysRevLett.50.1395 (18 May 1983).

108. Sorella, S. Green Function Monte Carlo with Stochastic Reconfiguration. *Phys. Rev. Lett.* **80,** 4558–4561. https://link.aps.org/doi/10.1103/PhysRevLett.80.4558 (20 May 1998).

109. Nomura, Y. *Boltzmann machines and quantum many-body problems* 2023. arXiv: 2306.16877 [cond-mat.str-el].

110. Nomura, Y., Darmawan, A., Yamaji, Y. & Imada, M. Restricted Boltzmann machine learning for solving strongly correlated quantum systems. *Phys. Rev. B* **96,** 205152. https://link.aps.org/doi/10.1103/PhysRevB.96.205152 (20 Nov. 2017).

111. Ferrari, F., Becca, F. & Carrasquilla, J. Neural Gutzwiller-projected variational wave functions. *Phys. Rev. B* **100,** 125131. https://link.aps.org/doi/10.1103/PhysRevB.100.125131 (12 Sept. 2019).

112. Cha, P. *et al.* Attention-based quantum tomography. *Machine Learning: Science and Technology* **3,** 01LT01. https://dx.doi.org/10.1088/2632-2153/ac362b (Nov. 2021).

113. Luo, D., Chen, Z., Carrasquilla, J. & Clark, B. Autoregressive Neural Network for Simulating Open Quantum Systems via a Probabilistic Formulation. *Phys. Rev. Lett.* **128,** 090501. https://link.aps.org/doi/10.1103/PhysRevLett.128.090501 (9 Feb. 2022).

114. White, S. & Affleck, I. Dimerization and incommensurate spiral spin correlations in the zigzag spin chain: Analogies to the Kondo lattice. *Phys. Rev. B* **54,** 9862–9869. https://link.aps.org/doi/10.1103/PhysRevB.54.9862 (14 Oct. 1996).

115. Eggert, S. Numerical evidence for multiplicative logarithmic corrections from marginal operators. *Phys. Rev. B* **54,** R9612–R9615. https://link.aps.org/doi/10.1103/PhysRevB.54.R9612 (14 Oct. 1996).

116. Sandvik, A. Computational Studies of Quantum Spin Systems. *AIP Conference Proceedings* **1297,** 135–338. https://aip.scitation.org/doi/abs/10.1063/1.3518900 (2010).

117. White, S. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69,** 2863–2866. https://link.aps.org/doi/10.1103/PhysRevLett.69.2863 (19 Nov. 1992).

118. Schollwöck, U. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics* **326,** 96–192. https://doi.org/10.1016/j.aop.2010.09.012 (1 Jan. 2011).

119. Becca, F. & Sorella, S. *Quantum Monte Carlo Approaches for Correlated Systems* https://www.cambridge.org/core/books/quantum-monte-carlo-approaches-for-correlated-systems/EB88C86BD9553A0738BDAE400D0B2900 (Cambridge University Press, 2017).

120. Marshall, W. Antiferromagnetism. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **232,** 48–68. http://www.jstor.org/stable/99682 (1955).

121. Ferrari, F., Parola, A., Sorella, S. & Becca, F. Dynamical structure factor of the $J_1 - J_2$ Heisenberg model in one dimension: The variational Monte Carlo approach. *Phys. Rev. B* **97,** 235103. https://link.aps.org/doi/10.1103/PhysRevB.97.235103 (23 June 2018).

122. Viteritti, L., Ferrari, F. & Becca, F. Accuracy of restricted Boltzmann machines for the one-dimensional $J_1 - J_2$ Heisenberg model. *SciPost Phys.* **12,** 166. https://scipost.org/10.21468/SciPostPhys.12.5.166 (2022).

123. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* **5,** 157–166 (1994).

124. Pippan, P., White, S. & Evertz, H. Efficient matrix-product state method for periodic boundary conditions. *Phys. Rev. B* **81,** 081103. https://link.aps.org/doi/10.1103/PhysRevB.81.081103 (8 Feb. 2010).

125. Shaw, P., Uszkoreit, J. & Vaswani, A. *Self-Attention with Relative Position Representations* Apr. 2018. http://arxiv.org/abs/1803.02155.

126. Westerhout, T., Astrakhantsev, N., Tikhonov, K., Katsnelson, M. & Bagrov, A. Neural Quantum States of frustrated magnets: generalization and sign structure. *Nature Communications* **11,** 1593. http://arxiv.org/abs/1907.08186 (Dec. 2020).

127. Szabó, A. & Castelnovo, C. Neural network wave functions and the sign problem. *Phys. Rev. Research* **2,** 033075. https://link.aps.org/doi/10.1103/PhysRevResearch.2.033075 (3 July 2020).

128. Park, C.-Y. & Kastoryano, M. *Expressive power of complex-valued restricted Boltzmann machines for solving non-stoquastic Hamiltonians* Aug. 2021. http://arxiv.org/abs/2012.08889.

129. Bukov, M., Schmitt, M. & Dupont, M. Learning the ground state of a non-stoquastic quantum Hamiltonian in a rugged neural network landscape. *SciPost Phys.* **10,** 147. https://scipost.org/10.21468/SciPostPhys.10.6.147 (2021).

130. Capriotti, L., Becca, F., Parola, A. & Sorella, S. Suppression of dimer correlations in the two-dimensional $J_1 - J_2$ Heisenberg model: An exact diagonalization study. *Phys. Rev. B* **67,** 212402. https://link.aps.org/doi/10.1103/PhysRevB.67.212402 (21 June 2003).

131. Bengio, Y., Courville, A. & Vincent, P. *Representation Learning: A Review and New Perspectives* 2014. arXiv: 1206.5538 [cs.LG].

132. LeCun, Y., Bengio, Y. & Hinton, G. Deep Learning. *Nature* **521,** 436–44 (May 2015).

133. Li, M. *et al.* Bridging the Gap between Deep Learning and Frustrated Quantum Spin System for Extreme-Scale Simulations on New Generation of Sunway Supercomputer. *IEEE Transactions on Parallel amp; Distributed Systems* **33,** 2846–2859. ISSN: 1558-2183 (Nov. 2022).

134. Liang, X. *et al.* 2^{1296} *Exponentially Complex Quantum Many-Body Simulation via Scalable Deep Learning Method* 2022. arXiv: 2204.07816 [quant-ph].

135. Xiong, R. *et al. On Layer Normalization in the Transformer Architecture* 2020. arXiv: 2002.04745 [cs.LG].

136. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* 2015. arXiv: 1512.03385 [cs.CV].

137. Agarap, A. F. *Deep Learning using Rectified Linear Units (ReLU)* 2019. arXiv: 1803.08375 [cs.NE].

138. Sorella, S. Wave function optimization in the variational Monte Carlo method. *Phys. Rev. B* **71,** 241103. https://link.aps.org/doi/10.1103/PhysRevB.71.241103 (24 June 2005).

139. Rende, R. & Viteritti, L. L. Are queries and keys always relevant? A case study on transformer wave functions. *Machine Learning: Science and Technology* **6,** 010501. https://dx.doi.org/10.1088/2632-2153/ada1a0 (Jan. 2025).

140. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* 2020. arXiv: 1802.03426 [stat.ML].

141. Rende, R., Goldt, S., Becca, F. & Viteritti, L. L. Fine-tuning neural network quantum states. *Phys. Rev. Res.* **6,** 043280. https://link.aps.org/doi/10.1103/PhysRevResearch.6.043280 (4 Dec. 2024).

142. Henderson, H. V. & Searle, S. R. On Deriving the Inverse of a Sum of Matrices. *Siam Review* **23,** 53–60. https://api.semanticscholar.org/CorpusID:123113511 (1981).

143. Petersen, K. B. & Pedersen, M. S. *The Matrix Cookbook* Nov. 2012. http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html.

144. Novak, R., Sohl-Dickstein, J. & Schoenholz, S. S. *Fast Finite Width Neural Tangent Kernel* in *Proceedings of the 39th International Conference on Machine Learning* **162** (PMLR, July 2022), 17018–17044. https://proceedings.mlr.press/v162/novak22a.html.

145. Vicentini, F. *et al.* NetKet 3: Machine Learning Toolbox for Many-Body Quantum Systems. *SciPost Phys. Codebases,* 7. https://scipost.org/10.21468/SciPostPhysCodeb.7 (2022).

146. Chen, A. & Heyl, M. *Efficient optimization of deep neural quantum states toward machine precision* 2023. arXiv: 2302.01941 [cond-mat.dis-nn].

147. Ledinauskas, E. & Anisimovas, E. *Scalable Imaginary Time Evolution with Neural Network Quantum States* 2023. arXiv: 2307.15521 [quant-ph].

148. Szabó, A. & Castelnovo, C. Neural network wave functions and the sign problem. *Phys. Rev. Res.* **2,** 033075. https://link.aps.org/doi/10.1103/PhysRevResearch.2.033075 (3 July 2020).

149. Liang, X. *et al.* Solving frustrated quantum many-particle models with convolutional neural networks. *Phys. Rev. B* **98,** 104426. https://link.aps.org/doi/10.1103/PhysRevB.98.104426 (10 Sept. 2018).

150. Choo, K., Neupert, T. & Carleo, G. Two-dimensional frustrated J_1-J_2 model studied with neural network quantum states. *Phys. Rev. B* **100,** 125124. https://link.aps.org/doi/10.1103/PhysRevB.100.125124 (12 Sept. 2019).

151. Liang, X., Dong, S.-J. & He, L. Hybrid convolutional neural network and projected entangled pair states wave functions for quantum many-particle states. *Phys. Rev. B* **103,** 035138. https://link.aps.org/doi/10.1103/PhysRevB.103.035138 (3 Jan. 2021).

152. Gong, S.-S., Zhu, W., Sheng, D. N., Motrunich, O. I. & Fisher, M. P. A. 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. https://link.aps.org/doi/10.1103/PhysRevLett.113.027201 (2 July 2014).

153. Wang, J.-Q., He, R.-Q. & Lu, Z.-Y. *Variational optimization of the amplitude of neural-network quantum many-body ground states* 2023. arXiv: 2308.09664 [cond-mat.str-el].

154. Ferrari, F., Becca, F. & Carrasquilla, J. Neural Gutzwiller-projected variational wave functions. *Phys. Rev. B* **100,** 125131. https://link.aps.org/doi/10.1103/PhysRevB.100.125131 (12 Sept. 2019).

155. Reh, M., Schmitt, M. & Gärttner, M. Optimizing design choices for neural quantum states. *Phys. Rev. B* **107,** 195115. https://link.aps.org/doi/10.1103/PhysRevB.107.195115 (19 May 2023).

156. Chen, H., Hendry, D., Weinberg, P. & Feiguin, A. *Systematic improvement of neural network quantum states using Lanczos* in *Advances in Neural Information Processing Systems* (eds Koyejo, S. *et al.*) **35** (Curran Associates, Inc., 2022), 7490–7503. https://proceedings.neurips.cc/paper_files/paper/2022/file/3173c427cb4ed2d5eaab029c17f221ae-Paper-Conference.pdf.

157. Li, M. *et al.* Bridging the Gap between Deep Learning and Frustrated Quantum Spin System for Extreme-Scale Simulations on New Generation of Sunway Supercomputer. *IEEE Transactions on Parallel amp; Distributed Systems* **33,** 2846–2859. ISSN: 1558-2183 (Nov. 2022).

158. Roth, C., Szabó, A. & MacDonald, A. H. High-accuracy variational Monte Carlo for frustrated magnets with deep neural networks. *Phys. Rev. B* **108,** 054410. https://link.aps.org/doi/10.1103/PhysRevB.108.054410 (5 Aug. 2023).

159. Liang, X. *et al.* 2^{1296} *Exponentially Complex Quantum Many-Body Simulation via Scalable Deep Learning Method* 2022. arXiv: 2204.07816 [quant-ph].

160. Hu, W.-J., Becca, F., Parola, A. & Sorella, S. Direct evidence for a gapless Z_2 spin liquid by frustrating Néel antiferromagnetism. *Phys. Rev. B* **88,** 060402. https://link.aps.org/doi/10.1103/PhysRevB.88.060402 (6 Aug. 2013).

161. Nomura, Y. & Imada, M. 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. https://link.aps.org/doi/10.1103/PhysRevX.11.031034 (3 Aug. 2021).

162. Nomura, Y. Helping restricted Boltzmann machines with quantum-state representation by restoring symmetry. *Journal of Physics: Condensed Matter* **33,** 174003. https://dx.doi.org/10.1088/1361-648X/abe268 (Apr. 2021).

163. Tsui, D. C., Stormer, H. L. & Gossard, A. C. Two-Dimensional Magnetotransport in the Extreme Quantum Limit. *Phys. Rev. Lett.* **48,** 1559–1562. https://link.aps.org/doi/10.1103/PhysRevLett.48.1559 (22 May 1982).

164. Savary, L. & Balents, L. Quantum spin liquids: a review. *Rep. Prog. Phys.* **80,** 016502. https://doi.org/10.1088/0034-4885/80/1/016502 (Nov. 2017).

165. Kitaev, A. Anyons in an exactly solved model and beyond. *Ann. Phys.* **321.** January Special Issue, 2–111. https://www.sciencedirect.com/science/article/pii/S0003491605002381 (2006).

166. Norman, M. Colloquium: Herbertsmithite and the search for the quantum spin liquid. *Rev. Mod. Phys.* **88,** 041002. https://link.aps.org/doi/10.1103/RevModPhys.88.041002 (4 Dec. 2016).

167. Onizuka, K. *et al.* 1/3 Magnetization Plateau in $SrCu_2(BO_3)_2$ - Stripe Order of Excited Triplets -. *Journal of the Physical Society of Japan* **69,** 1016–1018. https://doi.org/10.1143/JPSJ.69.1016 (2000).

168. Kodama, K. *et al.* Magnetic Superstructure in the Two-Dimensional Quantum Antiferromagnet $SrCu_2(BO_3)_2$. *Science* **298,** 395–399. https://www.science.org/doi/abs/10.1126/science.1075045 (2002).

169. Corboz, P. & Mila, F. Crystals of Bound States in the Magnetization Plateaus of the Shastry-Sutherland Model. *Phys. Rev. Lett.* **112,** 147203. https://link.aps.org/doi/10.1103/PhysRevLett.112.147203 (14 Apr. 2014).

170. Albrecht, M. & Mila, F. First-order transition between magnetic order and valence bond order in a 2D frustrated Heisenberg model. *Europhysics Letters* **34,** 145. https://dx.doi.org/10.1209/epl/i1996-00430-0 (Apr. 1996).

171. Weihong, Z., Hamer, C. J. & Oitmaa, J. Series expansions for a Heisenberg antiferromagnetic model for $SrCu_2(BO_3)_2$. *Phys. Rev. B* **60,** 6608–6616. https://link.aps.org/doi/10.1103/PhysRevB.60.6608 (9 Sept. 1999).

172. Koga, A. & Kawakami, N. Quantum Phase Transitions in the Shastry-Sutherland Model for $SrCu_2(BO_3)_2$. *Phys. Rev. Lett.* **84,** 4461–4464. https://link.aps.org/doi/10.1103/PhysRevLett.84.4461 (19 May 2000).

173. Chung, C. H., Marston, J. B. & Sachdev, S. Quantum phases of the Shastry-Sutherland antiferromagnet: Application to $SrCu_2(BO_3)_2$. *Phys. Rev. B* **64,** 134407. https://link.aps.org/doi/10.1103/PhysRevB.64.134407 (13 Sept. 2001).

174. Läuchli, A., Wessel, S. & Sigrist, M. Phase diagram of the quadrumerized Shastry-Sutherland model. *Phys. Rev. B* **66,** 014401. https://link.aps.org/doi/10.1103/PhysRevB.66.014401 (1 June 2002).

175. Corboz, P. & Mila, F. Tensor network study of the Shastry-Sutherland model in zero magnetic field. *Phys. Rev. B* **87,** 115144. https://link.aps.org/doi/10.1103/PhysRevB.87.115144 (11 Mar. 2013).

176. Waki, T. *et al.* A Novel Ordered Phase in SrCu2(BO3)2 under High Pressure. *Journal of the Physical Society of Japan* **76,** 073710. https://doi.org/10.1143/JPSJ.76.073710 (2007).

177. Zayed, M. E. *et al.* 4-spin plaquette singlet state in the Shastry–Sutherland compound $SrCu_2(BO_3)_2$. *Nature Physics* **13,** 962–966. https://doi.org/10.1038/nphys4190 (10 July 2017).

178. Guo, J. *et al.* Quantum Phases of $SrCu_2(BO_3)_2$ from High-Pressure Thermodynamics. *Phys. Rev. Lett.* **124,** 206602. https://link.aps.org/doi/10.1103/PhysRevLett.124.206602 (20 May 2020).

179. Lee, J. Y., You, Y.-Z., Sachdev, S. & Vishwanath, A. Signatures of a Deconfined Phase Transition on the Shastry-Sutherland Lattice: Applications to Quantum Critical $SrCu_2(BO_3)_2$. *Phys. Rev. X* **9,** 041037. https://link.aps.org/doi/10.1103/PhysRevX.9.041037 (4 Nov. 2019).

180. Liu, W.-Y. *et al. Deconfined quantum criticality with emergent symmetry in the extended Shastry-Sutherland model* 2023. arXiv: 2309.10955 [cond-mat.str-el].

181. Yang, J., Sandvik, A. W. & Wang, L. Quantum criticality and spin liquid phase in the Shastry-Sutherland model. *Phys. Rev. B* **105,** L060409. https://link.aps.org/doi/10.1103/PhysRevB.105.L060409 (6 Feb. 2022).

182. Wang, L., Zhang, Y. & Sandvik, A. W. Quantum Spin Liquid Phase in the Shastry-Sutherland Model Detected by an Improved Level Spectroscopic Method. *Chinese Physics Letters* **39,** 077502. https://dx.doi.org/10.1088/0256-307X/39/7/077502 (June 2022).

183. Kele ş, A. & Zhao, E. Rise and fall of plaquette order in the Shastry-Sutherland magnet revealed by pseudofermion functional renormalization group. *Phys. Rev. B* **105,** L041115. https://link.aps.org/doi/10.1103/PhysRevB.105.L041115 (4 Jan. 2022).

184. Fazekas, P. & Anderson, P. On the ground state properties of the anisotropic triangular antiferromagnet. *Phil. Mag.* **30,** 423–440 (1974).

185. Anderson, P. The resonating valence bond state in La_2CuO_4 and superconductivity. *Science* **235,** 1196–1198 (1987).

186. Ran, Y., Hermele, M., Lee, P. A. & Wen, X.-G. Projected-Wave-Function Study of the Spin-1/2 Heisenberg Model on the Kagomé Lattice. *Phys. Rev. Lett.* **98,** 117205. https://link.aps.org/doi/10.1103/PhysRevLett.98.117205 (11 Mar. 2007).

187. Iqbal, Y., Hu, W.-J., Thomale, R., Poilblanc, D. & Becca, F. Spin liquid nature in the Heisenberg $J_1 - J_2$ triangular antiferromagnet. *Phys. Rev. B* **93,** 144411. https://link.aps.org/doi/10.1103/PhysRevB.93.144411 (14 Apr. 2016).

188. Yan, S., Huse, D. A. & White, S. R. Spin-Liquid Ground State of the S=1/2 Kagome Heisenberg Antiferromagnet. *Science* **332,** 1173–1176. https://www.science.org/doi/abs/10.1126/science.1201080 (2011).

189. Zhu, Z. & White, S. R. Spin liquid phase of the $S = \frac{1}{2}$ $J_1 - J_2$ Heisenberg model on the triangular lattice. *Phys. Rev. B* **92,** 041105. https://link.aps.org/doi/10.1103/PhysRevB.92.041105 (4 July 2015).

190. Sandvik, A. W. Finite-size scaling of the ground-state parameters of the two-dimensional Heisenberg model. *Phys. Rev. B* **56,** 11678–11690. https://link.aps.org/doi/10.1103/PhysRevB.56.11678 (18 Nov. 1997).

191. Calandra Buonaura, M. & Sorella, S. 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,** 11446–11456. https://link.aps.org/doi/10.1103/PhysRevB.57.11446 (18 May 1998).

192. Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. What Does BERT Look At? An Analysis of BERT's Attention. *arXiv preprint arXiv:1906.04341.* arXiv: 1906.04341 [cs.CL] (2019).

193. Rogers, A., Kovaleva, O. & Rumshisky, A. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* **8,** 842–866. ISSN: 2307-387X. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00349/1923281/tacl_a_00349.pdf. https://doi.org/10.1162/tacl%5C_a%5C_00349 (Jan. 2021).

194. Jelassi, S., Sander, M. & Li, Y. *Vision Transformers provably learn spatial structure* in *Advances in Neural Information Processing Systems* (eds Koyejo, S. *et al.*) **35** (Curran Associates, Inc., 2022), 37822–37836. https://proceedings.neurips.cc/paper_files/paper/2022/file/f69707de866eb0805683d35217Paper-Conference.pdf.

195. Melko, R. G. & Carrasquilla, J. Language models for quantum simulation. *Nature Computat. Sci.* **4,** 11–18 (2024).

196. Sprague, K. & Czischek, S. *Variational Monte Carlo with Large Patched Transformers* 2023. arXiv: 2306.03921 [quant-ph].

197. Luo, D. *et al.* Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models. *Phys. Rev. Res.* **5,** 013216. https://link.aps.org/doi/10.1103/PhysRevResearch.5.013216 (1 Mar. 2023).

198. Von Glehn, I., Spencer, J. S. & Pfau, D. A Self-Attention Ansatz for Ab-initio Quantum Chemistry. *arXiv preprint arXiv:2211.13672.* arXiv: 2211.13672 [physics.chem-ph] (2023).

199. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683.* arXiv: 1910.10683 [cs.LG] (2023).

200. Dai, Z., Liu, H., Le, Q. V. & Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems* **34,** 3965–3977 (2021).

201. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-Attention with Relative Position Representations. *arXiv preprint arXiv:1803.02155.* arXiv: 1803.02155 [cs.CL] (2018).

202. Wennberg, U. & Henter, G. E. The Case for Translation-Invariant Self-Attention in Transformer-Based Language Models. *arXiv preprint arXiv:2106.01950.* arXiv: 2106.01950 [cs.CL] (2021).

203. Ke, G., He, D. & Liu, T.-Y. *Rethinking Positional Encoding in Language Pre-training* in *International Conference on Learning Representations* (2021). https://openreview.net/forum?id=09-528y2Fgf.

204. Gong, S.-S., Zhu, W., Sheng, D. N., Motrunich, O. I. & Fisher, M. P. A. 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. https://link.aps.org/doi/10.1103/PhysRevLett.113.027201 (2 July 2014).

205. Schulz, H. J., Ziman, T. A. & Poilblanc, D. Magnetic Order and Disorder in the Frustrated Quantum Heisenberg Antiferromagnet in Two Dimensions. *Journal de Physique I* **6,** 675–703. ISSN: 1286-4862. http://dx.doi.org/10.1051/jp1:1996236 (May 1996).

206. Nomura, Y. Helping restricted Boltzmann machines with quantum-state representation by restoring symmetry. *Journal of Physics: Condensed Matter* **33,** 174003. https://dx.doi.org/10.1088/1361-648X/abe268 (Apr. 2021).

207. Reh, M., Schmitt, M. & Gärttner, M. Optimizing design choices for neural quantum states. *Phys. Rev. B* **107,** 195115. https://link.aps.org/doi/10.1103/PhysRevB.107.195115 (19 May 2023).

208. Zayed, M. E. *et al.* 4-spin plaquette singlet state in the Shastry–Sutherland compound SrCu2(BO3)2. *Nature Physics* **13,** 962–966. ISSN: 1745-2481. http://dx.doi.org/10.1038/nphys4190 (July 2017).

209. Wichmann, E. H. & Crichton, J. H. Cluster Decomposition Properties of the S Matrix. *Phys. Rev.* **132,** 2788–2799. https://link.aps.org/doi/10.1103/PhysRev.132.2788 (6 Dec. 1963).

210. Weinberg, S. in *Conceptual Foundations of Quantum Field Theory* (ed Cao, T. Y.) 241–251 (Cambridge University Press, 1999).

211. Bommasani, R. *et al. On the Opportunities and Risks of Foundation Models* 2022. arXiv: 2108.07258 [cs.LG]. https://arxiv.org/abs/2108.07258.

212. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J., Doran, C. & Solorio, T.) (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 4171–4186. https://aclanthology.org/N19-1423.

213. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

214. Dosovitskiy, A. *et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* 2021. arXiv: 2010.11929 [cs.CV]. https://arxiv.org/abs/2010.11929.

215. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118,** e2016239118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2016239118. https://www.pnas.org/doi/abs/10.1073/pnas.2016239118 (2021).

216. Jumper, J. M. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

217. Brown, T. *et al. Language Models are Few-Shot Learners* in *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) **33** (Curran Associates, Inc., 2020), 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

218. Luo, D. & Clark, B. K. Backflow Transformations via Neural Networks for Quantum Many-Body Wave Functions. *Physical Review Letters* **122.** ISSN: 1079-7114. http://dx.doi.org/10.1103/PhysRevLett.122.226401 (June 2019).

219. Chen, A. & Heyl, M. Empowering deep neural quantum states through efficient optimization. *Nature Physics* **20,** 1476–1481 (2024).

220. Sprague, K. & Czischek, S. Variational Monte Carlo with large patched transformers. *Communications Physics* **7,** 90 (2024).

221. Zhang, Y.-H. & Di Ventra, M. Transformer quantum state: A multipurpose model for quantum many-body problems. *Phys. Rev. B* **107,** 075147. https://link.aps.org/doi/10.1103/PhysRevB.107.075147 (7 Feb. 2023).

222. Scherbela, M., Reisenhofer, R., Gerard, L., Marquetand, P. & Grohs, P. Solving the electronic Schrödinger equation for multiple nuclear geometries with weight-sharing deep neural networks. *Nature Computational Science* **2,** 1–11 (May 2022).

223. Gao, N. & Günnemann, S. *Ab-Initio Potential Energy Surfaces by Pairing GNNs with Neural Wave Functions* in *International Conference on Learning Representations* (2022).

224. Gao, N. & Günnemann, S. *Generalizing Neural Wave Functions* in *International Conference on Machine Learning (ICML)* (2023).

225. Miao, J., Hsieh, C.-Y. & Zhang, S.-X. Neural-network-encoded variational quantum algorithms. *Phys. Rev. Appl.* **21,** 014053. https://link.aps.org/doi/10.1103/PhysRevApplied.21.014053 (1 Jan. 2024).

226. Campos Venuti, L. & Zanardi, P. Quantum Critical Scaling of the Geometric Tensors. *Phys. Rev. Lett.* **99,** 095701. https://link.aps.org/doi/10.1103/PhysRevLett.99.095701 (9 Aug. 2007).

227. Zanardi, P., Giorda, P. & Cozzini, M. Information-Theoretic Differential Geometry of Quantum Phase Transitions. *Phys. Rev. Lett.* **99,** 100603. https://link.aps.org/doi/10.1103/PhysRevLett.99.100603 (10 Sept. 2007).

228. Yun, S., Jeong, M., Kim, R., Kang, J. & Kim, H. J. *Graph Transformer Networks* in *Advances in Neural Information Processing Systems* (eds Wallach, H. *et al.*) **32** (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/9d63484abb477c97640154d40595a3bb-Paper.pdf.

229. GU, S.-J. FIDELITY APPROACH TO QUANTUM PHASE TRANSITIONS. *International Journal of Modern Physics B* **24,** 4371–4458. ISSN: 1793-6578. http://dx.doi.org/10.1142/S0217979210056335 (Sept. 2010).

230. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nature Physics* **13,** 431–434. ISSN: 1745-2481. http://dx.doi.org/10.1038/nphys4035 (Feb. 2017).

231. Van Nieuwenburg, E. P. L., Liu, Y.-H. & Huber, S. D. Learning phase transitions by confusion. *Nature Physics* **13,** 435–439. ISSN: 1745-2481. http://dx.doi.org/10.1038/nphys4037 (Feb. 2017).

232. Huang, H.-Y., Kueng, R. & Preskill, J. Predicting many properties of a quantum system from very few measurements. *Nature Physics* **16,** 1050–1057. ISSN: 1745-2481. http://dx.doi.org/10.1038/s41567-020-0932-7 (June 2020).

233. Huang, H.-Y., Kueng, R., Torlai, G., Albert, V. V. & Preskill, J. Provably efficient machine learning for quantum many-body problems. *Science* **377.** ISSN: 1095-9203. http://dx.doi.org/10.1126/science.abk3333 (Sept. 2022).

234. Sachdev, S. Quantum phase transitions. *Physics world* **12,** 33 (1999).

235. Damski, B. Fidelity susceptibility of the quantum Ising model in a transverse field: The exact solution. *Phys. Rev. E* **87,** 052131. https://link.aps.org/doi/10.1103/PhysRevE.87.052131 (5 May 2013).

236. Schwandt, D., Alet, F. & Capponi, S. Quantum Monte Carlo Simulations of Fidelity at Magnetic Quantum Phase Transitions. *Phys. Rev. Lett.* **103,** 170501. https://link.aps.org/doi/10.1103/PhysRevLett.103.170501 (17 Oct. 2009).

237. Albuquerque, A. F., Alet, F., Sire, C. & Capponi, S. Quantum critical scaling of fidelity susceptibility. *Phys. Rev. B* **81,** 064418. https://link.aps.org/doi/10.1103/PhysRevB.81.064418 (6 Feb. 2010).

238. Binney, J. J., Dowrick, N. J., Fisher, A. J. & Newman, M. *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* ISBN: 0198513933 (Oxford University Press, Inc., USA, 1992).

239. Gelfand, M. P., Singh, R. R. P. & Huse, D. A. Zero-temperature ordering in two-dimensional frustrated quantum Heisenberg antiferromagnets. *Phys. Rev. B* **40,** 10801–10809. https://link.aps.org/doi/10.1103/PhysRevB.40.10801 (16 Dec. 1989).

240. Moreo, A., Dagotto, E., Jolicoeur, T. & Riera, J. Incommensurate correlations in the t-J and frustrated spin-1/2 Heisenberg models. *Phys. Rev. B* **42,** 6283–6293. https://link.aps.org/doi/10.1103/PhysRevB.42.6283 (10 Oct. 1990).

241. Chubukov, A. First-order transition in frustrated quantum antiferromagnets. *Phys. Rev. B* **44,** 392–394. https://link.aps.org/doi/10.1103/PhysRevB.44.392 (1 July 1991).

242. Read, N. & Sachdev, S. Large-N expansion for frustrated quantum antiferromagnets. *Phys. Rev. Lett.* **66,** 1773–1776. https://link.aps.org/doi/10.1103/PhysRevLett.66.1773 (13 Apr. 1991).

243. Mambrini, M., Läuchli, A., Poilblanc, D. & Mila, F. Plaquette valence-bond crystal in the frustrated Heisenberg quantum antiferromagnet on the square lattice. *Phys. Rev. B* **74,** 144422. https://link.aps.org/doi/10.1103/PhysRevB.74.144422 (14 Oct. 2006).

244. Sindzingre, P., Shannon, N. & Momoi, T. Phase diagram of the spin-1/2 J1-J2-J3 Heisenberg model on the square lattice. *Journal of Physics: Conference Series* **200,** 022058. https://dx.doi.org/10.1088/1742-6596/200/2/022058 (Jan. 2010).

245. Liu, W.-Y. *et al.* Emergence of Gapless Quantum Spin Liquid from Deconfined Quantum Critical Point. *Phys. Rev. X* **12,** 031039. https://link.aps.org/doi/10.1103/PhysRevX.12.031039 (3 Sept. 2022).

246. Liu, W.-Y., Poilblanc, D., Gong, S.-S., Chen, W.-Q. & Gu, Z.-C. Tensor network study of the spin-$\frac{1}{2}$ square-lattice $J_1{-}J_2{-}J_3$ model: Incommensurate spiral order, mixed valence-bond solids, and multicritical points. *Phys. Rev. B* **109,** 235116. https://link.aps.org/doi/10.1103/PhysRevB.109.235116 (23 June 2024).

247. Lacroix, C., Mendels, P. & Mila, F. Introduction to Frustrated Magnetism. *Introduction to Frustrated Magnetism: Materials, Experiments, Theory, Springer Series in Solid-State Sciences, Volume 164. ISBN 978-3-642-10588-3. Springer-Verlag Berlin Heidelberg, 2011* **-1** (Jan. 2011).

248. Parisi, G. Nobel lecture: Multiple equilibria. *Reviews of Modern Physics* **95,** 030501 (2023).

249. Abanin, D. A., Altman, E., Bloch, I. & Serbyn, M. Colloquium: Many-body localization, thermalization, and entanglement. *Reviews of Modern Physics* **91,** 021001 (2019).

250. Sierant, P., Lewenstein, M., Scardicchio, A., Vidmar, L. & Zakrzewski, J. Many-body localization in the age of classical computing. *Reports on Progress in Physics* **88,** 026502 (2025).

251. Viteritti, L. L. *et al. Quantum Spin Glass in the Two-Dimensional Disordered Heisenberg Model via Foundation Neural-Network Quantum States* 2025. arXiv: 2507.05073 [cond-mat.dis-nn]. https://arxiv.org/abs/2507.05073.

252. McCoy, B. M. & Wu, T. T. Theory of a two-dimensional Ising model with random impurities. I. Thermodynamics. *Physical Review* **176,** 631 (1968).

253. Fisher, D. S. Random transverse field Ising spin chains. *Phys. Rev. Lett.* **69,** 534–537. https://link.aps.org/doi/10.1103/PhysRevLett.69.534 (3 July 1992).

254. Young, A. P. & Rieger, H. Numerical study of the random transverse-field Ising spin chain. *Phys. Rev. B* **53,** 8486–8498. https://link.aps.org/doi/10.1103/PhysRevB.53.8486 (13 Apr. 1996).

255. Krämer, C., Koziol, J. A., Langheld, A., Hörmann, M. & Schmidt, K. P. Quantum-critical properties of the one- and two-dimensional random transverse-field Ising model from large-scale quantum Monte Carlo simulations. *SciPost Physics* **17.** ISSN: 2542-4653. http://dx.doi.org/10.21468/SciPostPhys.17.2.061 (Aug. 2024).

256. Coldea, R., Tennant, D. A., Tsvelik, A. M. & Tylczynski, Z. Experimental Realization of a 2D Fractional Quantum Spin Liquid. *Phys. Rev. Lett.* **86,** 1335–1338. https://link.aps.org/doi/10.1103/PhysRevLett.86.1335 (7 Feb. 2001).

257. Heidarian, D., Sorella, S. & Becca, F. Spin-$\frac{1}{2}$ Heisenberg model on the anisotropic triangular lattice: From magnetism to a one-dimensional spin liquid. *Phys. Rev. B* **80,** 012404. https://link.aps.org/doi/10.1103/PhysRevB.80.012404 (1 July 2009).

258. Hasik, J. & Corboz, P. Incommensurate Order with Translationally Invariant Projected Entangled-Pair States: Spiral States and Quantum Spin Liquid on the Anisotropic Triangular Lattice. *Phys. Rev. Lett.* **133,** 176502. https://link.aps.org/doi/10.1103/PhysRevLett.133.176502 (17 Oct. 2024).

259. Wolf, T. *et al. Transformers: State-of-the-Art Natural Language Processing* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics, Online, Oct. 2020), 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6.

260. Liu, Z. *et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows* 2021. arXiv: 2103.14030 [cs.CV]. https://arxiv.org/abs/2103.14030.

261. Scalapino, D. J. A common thread: The pairing interaction for unconventional superconductors. *Rev. Mod. Phys.* **84,** 1383–1417. https://link.aps.org/doi/10.1103/RevModPhys.84.1383 (4 Oct. 2012).

262. Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: Physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78,** 17–85. https://link.aps.org/doi/10.1103/RevModPhys.78.17 (1 Jan. 2006).

263. Corboz, P., Rice, T. M. & Troyer, M. Competing States in the t-J Model: Uniform d-Wave State versus Stripe State. *Phys. Rev. Lett.* **113,** 046402. https://link.aps.org/doi/10.1103/PhysRevLett.113.046402 (4 July 2014).

264. Zheng, B.-X. *et al.* Stripe order in the underdoped region of the two-dimensional Hubbard model. *Science* **358,** 1155–1160. eprint: https://www.science.org/doi/pdf/10.1126/science.aam7127. https://www.science.org/doi/abs/10.1126/science.aam7127 (2017).

265. Luo, D. & Clark, B. K. Backflow Transformations via Neural Networks for Quantum Many-Body Wave Functions. *Phys. Rev. Lett.* **122,** 226401. https://link.aps.org/doi/10.1103/PhysRevLett.122.226401 (22 June 2019).

266. Choo, K., Mezzacapo, A. & Carleo, G. Fermionic neural-network states for ab-initio electronic structure. *Nature communications* **11,** 2368 (2020).

267. Hermann, J. *et al.* Ab initio quantum chemistry with neural-network wavefunctions. *Nature Reviews Chemistry* **7,** 692–709 (2023).

268. Scherbela, M., Gerard, L. & Grohs, P. Towards a transferable fermionic neural wavefunction for molecules. *Nature Communications* **15,** 120 (2024).

269. Sinibaldi, A., Hendry, D., Vicentini, F. & Carleo, G. *Time-dependent Neural Galerkin Method for Quantum Dynamics* 2024. arXiv: 2412.11778 [quant-ph]. https://arxiv.org/abs/2412.11778.

270. De Walle, A. V., Schmitt, M. & Bohrdt, A. *Many-body dynamics with explicitly time-dependent neural quantum states* 2024. arXiv: 2412.11830 [quant-ph]. https://arxiv.org/abs/2412.11830.

271. Acevedo, S., Rodriguez, A. & Laio, A. Unsupervised detection of semantic correlations in big data. *Communications Physics* **8.** ISSN: 2399-3650. http://dx.doi.org/10.1038/s42005-025-02115-z (May 2025).

272. Acevedo, S. *et al. An approach to identify the most semantically informative deep representations of text and images* 2025. arXiv: 2505.17101 [cs.CL]. https://arxiv.org/abs/2505.17101.

273. Mendes-Santos, T., Turkeshi, X., Dalmonte, M. & Rodriguez, A. Unsupervised Learning Universal Critical Behavior via the Intrinsic Dimension. *Phys. Rev. X* **11,** 011040. https://link.aps.org/doi/10.1103/PhysRevX.11.011040 (1 Feb. 2021).

274. Wolf, T. *et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing* 2020. arXiv: 1910.03771 [cs.CL].

275. Izsak, P., Berchansky, M. & Levy, O. *How to Train BERT with an Academic Budget* 2021. arXiv: 2104.07705 [cs.CL].

276. Arnold, B. C. & Press, S. J. Compatible Conditional Distributions. *Journal of the American Statistical Association.* http://www.jstor.org/stable/2289858 (1989).

277. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. *BERTScore: Evaluating Text Generation with BERT* 2020. arXiv: 1904.09675 [cs.CL].

278. Xiong, R. *et al. On Layer Normalization in the Transformer Architecture* 2020. arXiv: 2002.04745 [cs.LG].

279. Reh, M., Schmitt, M. & Gärttner, M. Optimizing design choices for neural quantum states. *Phys. Rev. B* **107,** 195115. https://link.aps.org/doi/10.1103/PhysRevB.107.195115 (19 May 2023).

280. Yuan, X., Endo, S., Zhao, Q., Li, Y. & Benjamin, S. C. Theory of variational quantum simulation. *Quantum* **3,** 191. ISSN: 2521-327X. http://dx.doi.org/10.22331/q-2019-10-07-191 (Oct. 2019).

281. Schmitt, M. & Heyl, M. Quantum Many-Body Dynamics in Two Dimensions with Artificial Neural Networks. *Phys. Rev. Lett.* **125,** 100503. https://link.aps.org/doi/10.1103/PhysRevLett.125.100503 (10 Sept. 2020).

282. Stokes, J., Izaac, J., Killoran, N. & Carleo, G. Quantum Natural Gradient. *Quantum* **4,** 269. ISSN: 2521-327X. http://dx.doi.org/10.22331/q-2020-05-25-269 (May 2020).

283. Wu, D. *et al.* Variational benchmarks for quantum many-body problems. *Science* **386,** 296–301. eprint: https://www.science.org/doi/pdf/10.1126/science.adg9774. https://www.science.org/doi/abs/10.1126/science.adg9774 (2024).