

Convolutional transformer wave functions

Ao Chen^{1,2,*}, Vighnesh Dattatraya Naik^{1,*}, and Markus Heyl¹¹Center for Electronic Correlations and Magnetism, University of Augsburg, 86135 Augsburg, Germany²Center for Computational Quantum Physics, Flatiron Institute, New York, New York 10010, USA

(Received 30 April 2025; accepted 15 May 2026; published 3 June 2026)

Deep neural quantum states have recently achieved remarkable performance in solving challenging quantum many-body problems. While transformer networks appear particularly promising due to their success in computer science, we show that previously reported transformer wave functions have not so far been capable of utilizing their full power. Here, we introduce the convolutional transformer wave function (CTWF). We show that our CTWF exhibits state-of-the-art performance in ground-state search and nonequilibrium dynamics compared to previous results, demonstrating its promising capacity in complex quantum problems.

DOI: [10.1103/PhysRevResearch.8.L022040](https://doi.org/10.1103/PhysRevResearch.8.L022040)

Introduction. The accurate numerical solution of strongly correlated quantum matter remains an outstanding challenge in modern quantum physics. This concerns, in particular, the regime of large two-dimensional quantum many-body systems despite impressive theoretical developments. In recent years, the neural quantum state (NQS) has emerged as a promising numerical method to solve the quantum many-body problem. The NQS is based on utilizing artificial neural networks (ANNs) to encode the quantum many-body wave functions [1]. As ANNs are universal function approximations, the NQS becomes a numerically exact approach converging, in principle, to the exact solution upon increasing the size of the ANN. Until now, the NQS has shown great potential in various quantum many-body problems, including quantum spin liquids [2–4], Fermi-Hubbard models [5–7], electronic structures [8–12], open quantum systems [13–15], and quantum dynamics [16–21].

The key enabling potential for the NQS technique is the expressive power of the underlying ANNs, in particular, when it comes to modern deep architectures. While the initial starting points were based on still relatively shallow networks such as restricted Boltzmann machines (RBMs) [1,22], in the recent years many deeper networks have already been investigated, including convolutional neural networks (CNNs) [23], variational autoregressive networks [24], recurrent neural networks (RNNs) [25], group CNNs [26], and deep CNNs [4,27]. As a consequence of recent advances in training algorithms, it has now also become possible to optimize deep NQSs with up to 10^6 parameters, thereby pushing the NQS

approach more toward exploiting the full power of ANNs. In particular, with such deep NQSs unprecedented numerical accuracies have been reached for multiple frustrated quantum magnets [4].

In the field of machine learning, transformer neural networks have developed into the most powerful architecture for many tasks [28,29]. Consequently, it is a natural immediate question to what extent such transformer architectures could have a similar powerful potential for NQSs. So far, in the context of NQSs, transformers have already been applied in electronic structure problems [12,30–32] and quantum lattice models [33,34]. However, as we will discuss in this work, these previously introduced design choices in particular for lattice models are not in such a form yet so as to fully exploit the power of transformer architectures: They either are actually equivalent to CNNs, and therefore cannot fully exploit the key attention mechanism in transformers, or break translation symmetries. Thus, it has so far remained open how to fully utilize the full potential of transformer architectures in the context of quantum lattice models.

In this work, we introduce the convolutional transformer wave function (CTWF) inspired by recently developed variants of transformers in computer vision tasks [35,36]. By incorporating design choices rooted in physics principles, we tailor the CTWF for quantum spin systems and evaluate its performance by comparing against state-of-the-art results in the literature. We benchmark the CTWF's performance on two challenging problems: the ground-state search of the prototypical 10×10 J_1 – J_2 Heisenberg model, realizing a frustrated quantum magnet and quantum quench dynamics in the two-dimensional quantum Ising model. Our results demonstrate that the CTWF outperforms existing literature. In addition, we introduce a CNN architecture termed CNN (GELU), where GELU denotes the Gaussian Error Linear Unit, constituting an improved CNN that builds on our previous work [4]. The comparison to the CTWF yields a similar level of accuracy when using a similar number of parameters. We conclude that both the CTWF and CNN (GELU) achieve superior performance compared to previous methods, establishing state-of-the-art benchmarks for these complex quantum systems.

*These authors contributed equally to this work.

†Contact author: chenao.phys@gmail.com‡Contact author: naikvighnesh123@gmail.com

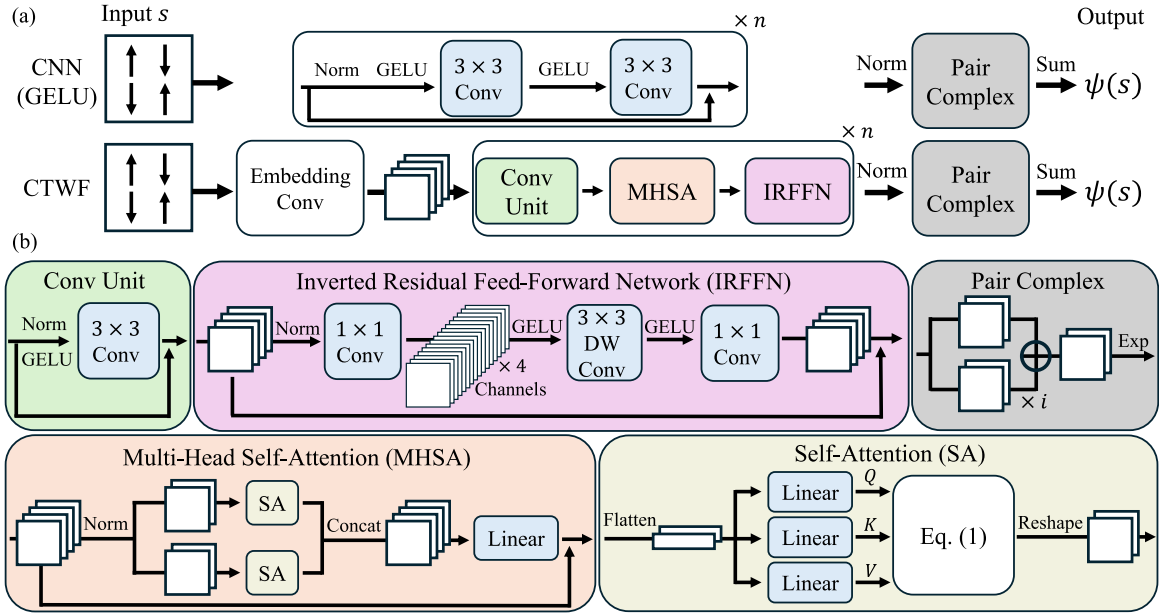


FIG. 1. (a) Network architecture of the CNN (GELU) improved from Ref. [4] and the CTWF. (b) Building blocks of CNN (GELU) and CTWF. The main block of CTWF consists of three consecutive parts, namely, the convolutional unit (Conv Unit), the MHSA, and the inverted residual feed-forward network (IRFFN). A normalization step (Norm) is applied at the beginning of all these parts by dividing the inputs by the expected initial standard deviation. The MHSA contains multiple self-attention (SA) blocks, each processed by Eq. (1). The IRFFN consists of three consecutive convolutional layers, and the middle one with an expanded channel dimension is depthwise (DW) to reduce the computing cost. Finally, the pair complex activation function is applied to convert real values to complex outputs [26]. The illustration shown here has channel dimension $c = 4$, self-attention token dimension $d = 2$, and the number of heads $h = 2$.

Transformer architectures. In the following, we revisit previously utilized transformer wave functions and introduce our CTWF whose structure is illustrated in Fig. 1. The core ingredient of transformer neural networks is the multihead self-attention (MHSA), which can be viewed as a parallelization of h attention heads [28]. Each attention head takes the same input $x \in \mathbb{R}^{n \times c}$, where $n = l_x \times l_y$ represents the flattened spatial dimension and c represents the embedded token dimension (or channel dimension in the context of CNN). The attention output is $A \in \mathbb{R}^{n \times d}$ given by $A_i = \sum_j \alpha(Q_i, K_j, P_{ij})V_j$, where $\alpha \in \mathbb{R}^{n \times n}$ is the attention coefficient, and $Q \in \mathbb{R}^{n \times d}$, $K \in \mathbb{R}^{n \times d}$, $V \in \mathbb{R}^{n \times d}$, and $P \in \mathbb{R}^{n \times n}$ are query, key, value, and positional encoding, respectively. Here, d is the so-called token dimension chosen as $d = c/h$ in this work. In the following, we discuss multiple design choices for the attention mechanism in the existing literature.

In the original transformer [28] and the vision transformer (ViT) [29], $Q = xW^Q$, $K = xW^K$, and $V = xW^V$, where $W^{Q/K/V} \in \mathbb{R}^{c \times d}$ are trainable parameters. $\alpha = \text{softmax}(QK^T/\sqrt{d})$, with $\text{softmax}(X)_{ij} = \exp(X_{ij})/\sum_{k=1}^N \exp(X_{ik})$. The full attention formula is $A = \text{softmax}(QK^T/\sqrt{d})V$, which is most popular in machine learning, recently also adopted in quantum lattice models with translation symmetry broken by positional encoding [37]. The central issue in this design choice is how to efficiently utilize the spatial information of the input while keeping translation symmetry.

In Refs. [33,38], the simplified ViT with factored attention is applied to quantum lattice models. In the factored attention, α is directly given by the relative positional encoding (RPE)

[39] $P_{ij} = p_{x_i - x_j, y_i - y_j}$, where $p \in \mathbb{R}^{l_x \times l_y}$ is trainable. The value V is still $V = xW^V$, so the full attention is $A = PxW^V$. In this attention mechanism, P is a linear transformation with a circular structure due to RPE, equivalent to a convolutional layer with full-size convolution kernel p . W^V can be viewed as a linear layer or a 1×1 convolution. Therefore, the factored attention represents essentially two consecutive convolutional layers instead of a normal attention layer, as also discussed in Ref. [40]. This factored attention has been adopted in several quantum many-body problems [41–43].

The autoregressive transformer quantum state is implemented by masking out the contribution from $i > j$ [34,44,45]. The attention is given by $A = \text{softmax}(QK^T/\sqrt{d} + M)V$, where $M_{ij} = 0$ for $i \leq j$ and $M_{ij} = -\infty$ for $i > j$. This attention mechanism allows uncorrelated sampling due to its autoregressive property but hence suffers from the strong constraint on architecture and the broken translation symmetry.

In summary, while previous works have already considered transformers, the power of transformer architectures in NQS for quantum many-body systems has not been fully harnessed in the existing literature. This motivates us to introduce the aforementioned CTWF based on the recent progress of transformers in computer vision [35,36]. The attention is designed by adding RPE to the original attention mechanism, i.e.,

$$A = \text{softmax}\left(\frac{QK^T + P}{\sqrt{d}}\right)V, \quad (1)$$

which implements an attention layer with translation symmetry. The attention outputs from different self-attention heads are concatenated after Eq. (1).

The whole network is designed according to the convolutional transformer architecture in computer vision [36] with some modifications. The attention layer is sandwiched by two convolutional blocks, namely, a convolutional unit and an IRFFN, to enhance its ability in representing local features and keep the translation symmetry across the whole network. We expect that the combination of MHSA and convolutional layers can help the network to encode both long-range and short-range correlations. An illustration of our network architecture is shown in Fig. 1.

In order to challenge the performance of the CTWF architecture, we also compare it to CNN (GELU), our most advanced CNN architecture to date, which we introduce here as an improvement over the CNN used in our previous work [4]. This improvement mainly stems from using the GELU activation [46] instead of the previously chosen Rectified Linear Unit to align with the choice in transformers, which yields higher accuracy in ground-state search and more stable time evolution in dynamics. The GELU activation is given by $\text{GELU}(x) = x \Phi(x)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

Numerical results. For the assessment of the performance of different network architectures in NQS ground-state search, we train them for the paradigmatic frustrated J_1 - J_2 Heisenberg model on a square lattice:

$$\mathcal{H} = J_1 \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \mathbf{S}_i \cdot \mathbf{S}_j, \quad (2)$$

where $\mathbf{S}_i = (S_i^x, S_i^y, S_i^z)$ denotes spin-1/2 operators at site i . $\langle i, j \rangle$ and $\langle\langle i, j \rangle\rangle$ indicate pairs of nearest and next-nearest neighbor sites, respectively. In this work, we specifically focus on the maximally frustrated point at $J_2/J_1 = 0.5$.

We assess the accuracy of the variational state $|\psi\rangle$ with respect to the exact ground state $|\phi\rangle$ through several quantities, including the relative error of energy $\epsilon_{\text{rel}} = (E - E_0)/|E_0|$, with $E = \langle\psi|\mathcal{H}|\psi\rangle/\langle\psi|\psi\rangle$ and $E_0 = \langle\phi|\mathcal{H}|\phi\rangle/\langle\phi|\phi\rangle$, the rescaled energy variance $\sigma^2/N = (\langle\mathcal{H}^2\rangle - E^2)/N$, where N denotes the system size, and the infidelity $I = 1 - \langle\psi|\phi\rangle\langle\phi|\psi\rangle/\langle\psi|\psi\rangle\langle\phi|\phi\rangle$, which is accessible in small systems. ϵ_{rel} , V , and I all tend to zero when the variational state $|\psi\rangle$ approaches the exact ground state $|\phi\rangle$.

We start by comparing different NQSs on the 6×6 lattice where the exact ground state $|\phi\rangle$ and the infidelity I are available. Each network is trained by stochastic reconfiguration [47] with 10^4 Monte Carlo samples for 10^4 steps. To ensure a fair comparison among different networks, we keep the number of network parameters N_p at a similar level to control the network size, and the number of multiply accumulate operations (MACs) at a similar level to control the computational cost. In this way, the comparison targets not only the accuracy of the ground-state solution but rather the performance under similar usage of computational resources.

The performance of RBM, CNN (GELU), and different transformers is shown in Table I. All deep networks studied in this work significantly outperform the shallow RBM, showing the necessity of modern deep NQSs for complex quantum systems. The factored attention, as we explained, is equivalent to

TABLE I. Performance comparison of various NQS architectures in the Heisenberg J_1 - J_2 model on the 6×6 square lattice. The tested networks include complex-valued RBM with translation symmetry, CNN (GELU), the transformer with factored attention [33,38,41], and CTWF. Bold indicates the best result in each quantity.

NQS	N_p	MACs	ϵ_{rel}	σ^2/N	I
RBM	9472	331776	0.0153	0.0228	0.1376
CNN	7120	254016	0.0024	0.0047	0.0227
Factored	7992	317920	0.0040	0.0083	0.0283
CTWF	7884	309420	0.0023	0.0055	0.0137

a CNN with full kernels. In Table I, its performance does not reach up to CTWF or CNN (GELU) with small kernels, which appears superior as also observed in recent works [4,26,27].

As a next step, we now challenge the performance of the CTWF for the 10×10 J_1 - J_2 Heisenberg model in Fig. 2. We choose CTWF with $n = 5$, $c = 48$, $d = 12$, $h = 4$, and $N_p = 255440$, and CNN (GELU) with $n = 8$ and $c = 32$ is also included. Apart from the inherent translation symmetry in these networks, we also apply symmetry projections, including spatial reflection, rotation, and spin flip, which amounts to 16 symmetry group elements in total. Given a similar amount of parameters, the variational energy of our CTWF and CNN (GELU) significantly outperforms the factored attention [40]. With more samples $N_s = 2^{14}$ and more parameters $N_p = 434760$, the factored attention is possible to reach variational energy $-0.4976764(7)$ and $\epsilon_{\text{rel}} = 7.9 \times 10^{-5}$ [48] compatible with the best previous CNN result, while still less accurate than CTWF. The result of autoregressive transformers is not shown due to the lack of reference, but here we present the result from an RNN [25] for illustration. The ϵ_{rel} and rescaled energy variance σ^2/N of this RNN

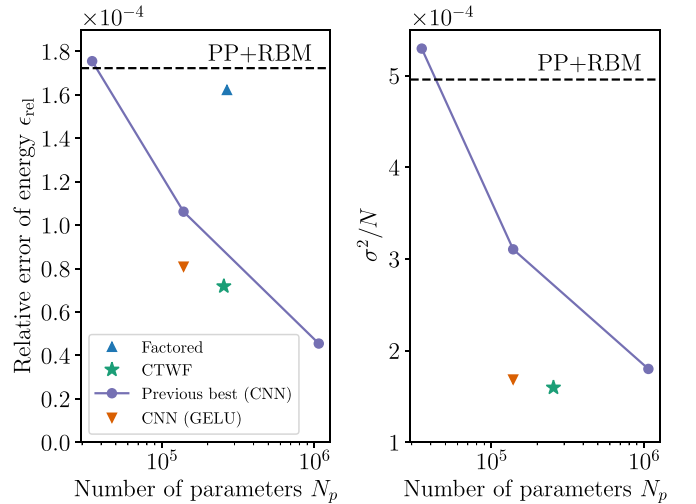


FIG. 2. The relative error of energy ϵ_{rel} and the rescaled energy variance σ^2/N in the 10×10 J_1 - J_2 Heisenberg model at $J_2/J_1 = 0.5$. The results presented here include the combination of the pair product state (PP) and RBM [2], the best previous results given by a CNN in Ref. [4], the transformer with factored attention [40], the CNN (GELU), and the CTWF. The reference ground-state energy is estimated by zero-variance extrapolation in Ref. [4].

are, respectively, 4.7×10^{-3} and 1.0×10^{-3} [49], which are significantly higher than the values in Fig. 2 and are hence not included in the figure. This suggests that the constrained network architecture and broken translation symmetry might be the key limiting factors in autoregressive NQSs including autoregressive transformers. Compared with the best previous result [4], CTWF and CNN (GELU) show similar accuracy in variational energy but produce a lower variance. These results show that these architectures lead to competitive variational wave functions for frustrated quantum magnets, which can serve as alternatives to existing CNN and transformer wave functions in the future.

Furthermore, we emphasize that the application of CTWF is not limited to ground-state searches. As an example, we present a simulation of real-time dynamics using CTWF. For this study, we employ the prototypical transverse-field Ising model,

$$\mathcal{H} = -J \sum_{(i,j)} \sigma_i^z \sigma_j^z - h \sum_i \sigma_i^x, \quad (3)$$

which has been widely used in various neural network quantum dynamics studies [16–18,50,51]. The simulation is conducted at the most challenging parameter point of the underlying quantum phase transition, $h/J = 3.044\,38(2)$ [52], on 6×6 and 8×8 square lattices. The quench dynamics begins from a paramagnetically polarized state, $|\psi_0\rangle = |\rightarrow\rangle^{\otimes N}$, and evolves under the Hamiltonian in Eq. (3). The real-time evolution is performed using the time-dependent variational principle (TDVP) with a fixed time interval of $J\tau = 10^{-3}$, employing the second-order Heun method.

We evaluate performance by comparing the expectation value $\langle \sigma_i^x(t) \rangle$ obtained using the CNN (GELU), the CTWF, and the best previous TDVP result (denoted as CNN-Schmitt) in Ref. [16] (see Fig. 3). For 6×6 , we also compare this with the exact result obtained by integrating the time-dependent Schrödinger equation. For both system sizes, our approach successfully extends the stability of the dynamics for a longer time using both CNN (GELU) and CTWF compared to previous benchmarks. This not only highlights the improved accuracy and robustness of CTWF and CNN (GELU) but also demonstrates that CTWF is a powerful neural quantum state for studying quantum dynamics, in addition to its well-established effectiveness in ground-state calculations.

Discussion. In this work, we have introduced the CTWF. We have found compelling evidence that this NQS architecture exhibits outstanding performance for both ground-state search and nonequilibrium quantum dynamics as compared to existing results in the literature. While these results highlight the potential of transformers for solving quantum lattice models, it is also important to emphasize that no definite conclusion on the superiority of transformers can be reached at this point, as an also introduced CNN (GELU) network structure yields comparable results. However, considering that the study of transformer wave functions is still at a comparatively early stage, it appears possible that attention-based architectures might outperform CNNs upon further developments. For instance, the transformer backflow has presented outstanding accuracy in a recent work [53].

One issue we face in this work is the computing cost of transformers due to self-attention, which originates from the

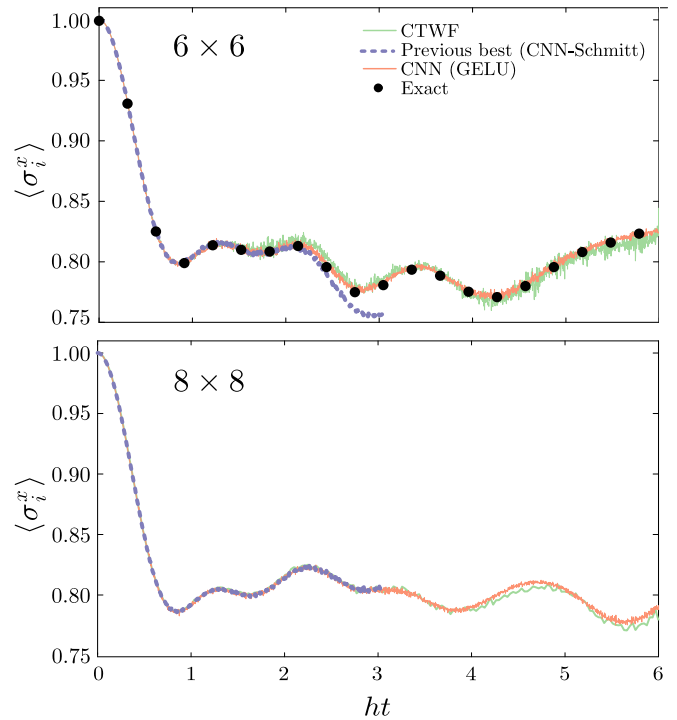


FIG. 3. Quench dynamics in the transverse-field Ising model at the critical point simulated by CTWF. The best previous TDVP result (CNN-Schmitt) [16] is shown for reference.

$\mathcal{O}(N^2)$ complexity of self-attention as compared to the $\mathcal{O}(N)$ complexity of CNN, where N is the system size. With a suitable choice of the embedding stride, nevertheless, the system size can be reduced to $\sqrt{N}$ to keep $\mathcal{O}(N)$ complexity. As transformers have been popular in the community of artificial intelligence, we expect the relevant progress in theory, software, and hardware can benefit our CTWF in the future and help us to scale CTWF to more parameters and larger systems. We therefore expect great potential for further improvements.

Acknowledgments. The exact results including ground state and real-time evolution are computed by the QUSPIN package [54]. The NQS ground-state search is simulated with Quantax [55], while the dynamics is simulated with jVMC [56]. We thank Markus Schmitt for providing the data for Fig. 3. We also acknowledge fruitful discussions with C. Roth, L. L. Viteritti, R. Rende, and F. Becca. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 853443). This work was supported by the German Research Foundation DFG via Project No. 492547816 (TRR 360). We gratefully acknowledge the scientific support and high-performance computing resources provided by the Scientific Computing Core at the Flatiron Institute, the LiCCA HPC cluster of the University of Augsburg, and the Erlangen National High Performance Computing Center (NHR) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (NHR Project No. nqsQuMat). The LiCCA cluster is cofunded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project No. 499211671. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially

funded by the German Research Foundation (Grant No. 440719683).

Data availability. The data that support the findings of this article are openly available [57]; embargo periods may apply.

- [1] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
- [2] Y. Nomura and M. Imada, 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 (2021).
- [3] N. Astrakhantsev, T. Westerhout, A. Tiwari, K. Choo, A. Chen, M. H. Fischer, G. Carleo, and T. Neupert, Broken-symmetry ground states of the Heisenberg model on the pyrochlore lattice, *Phys. Rev. X* **11**, 041021 (2021).
- [4] A. Chen and M. Heyl, Empowering deep neural quantum states through efficient optimization, *Nat. Phys.* **20**, 1476 (2024).
- [5] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Restricted Boltzmann machine learning for solving strongly correlated quantum systems, *Phys. Rev. B* **96**, 205152 (2017).
- [6] D. Luo and B. K. Clark, Backflow transformations via neural networks for quantum many-body wave functions, *Phys. Rev. Lett.* **122**, 226401 (2019).
- [7] J. R. Moreno, G. Carleo, A. Georges, and J. Stokes, Fermionic wave functions from neural-network constrained hidden states, *Proc. Natl. Acad. Sci. USA* **119**, e2122059119 (2022).
- [8] J. Hermann, J. Spencer, K. Choo, A. Mezzacapo, W. M. C. Foulkes, D. Pfau, G. Carleo, and F. Noé, *Ab initio* quantum chemistry with neural-network wavefunctions, *Nat. Rev. Chem.* **7**, 692 (2023).
- [9] K. Choo, A. Mezzacapo, and G. Carleo, Fermionic neural-network states for *ab-initio* electronic structure, *Nat. Commun.* **11**, 2368 (2020).
- [10] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, *Ab initio* solution of the many-electron Schrödinger equation with deep neural networks, *Phys. Rev. Res.* **2**, 033429 (2020).
- [11] J. Hermann, Z. Schätzle, and F. Noé, Deep-neural-network solution of the electronic Schrödinger equation, *Nat. Chem.* **12**, 891 (2020).
- [12] D. Pfau, S. Axelrod, H. Sutterud, I. von Glehn, and J. S. Spencer, Accurate computation of quantum excited states with neural networks, *Science* **385**, eadn0137 (2024).
- [13] A. Nagy and V. Savona, Variational quantum Monte Carlo method with a neural-network ansatz for open quantum systems, *Phys. Rev. Lett.* **122**, 250501 (2019).
- [14] M. J. Hartmann and G. Carleo, Neural-network approach to dissipative quantum many-body dynamics, *Phys. Rev. Lett.* **122**, 250502 (2019).
- [15] F. Vicentini, A. Biella, N. Regnault, and C. Ciuti, Variational neural-network ansatz for steady states in open quantum systems, *Phys. Rev. Lett.* **122**, 250503 (2019).
- [16] M. Schmitt and M. Heyl, Quantum many-body dynamics in two dimensions with artificial neural networks, *Phys. Rev. Lett.* **125**, 100503 (2020).
- [17] M. Schmitt, M. M. Rams, J. Dziarmaga, M. Heyl, and W. H. Zurek, Quantum phase transition dynamics in the two-dimensional transverse-field Ising model, *Sci. Adv.* **8**, eabl6850 (2022).
- [18] A. Sinibaldi, C. Giuliani, G. Carleo, and F. Vicentini, Unbiasing time-dependent variational Monte Carlo by projected quantum evolution, *Quantum* **7**, 1131 (2023).
- [19] T. Mendes-Santos, M. Schmitt, and M. Heyl, Highly resolved spectral functions of two-dimensional systems with neural quantum states, *Phys. Rev. Lett.* **131**, 046501 (2023).
- [20] J. Nys, G. Pescia, A. Sinibaldi, and G. Carleo, *Ab-initio* variational wave functions for the time-dependent many-electron Schrödinger equation, *Nat. Commun.* **15**, 9404 (2024).
- [21] T. Mendes-Santos, M. Schmitt, A. Angelone, A. Rodriguez, P. Scholl, H. J. Williams, D. Barredo, T. Lahaye, A. Browaeys, M. Heyl, and M. Dalmonte, Wave-function network description and Kolmogorov complexity of quantum many-body systems, *Phys. Rev. X* **14**, 021029 (2024).
- [22] Y. Nomura, Helping restricted Boltzmann machines with quantum-state representation by restoring symmetry, *J. Phys.: Condens. Matter* **33**, 174003 (2021).
- [23] K. Choo, T. Neupert, and G. Carleo, Two-dimensional frustrated J_1 - J_2 model studied with neural network quantum states, *Phys. Rev. B* **100**, 125124 (2019).
- [24] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, Deep autoregressive models for the efficient variational simulation of many-body quantum systems, *Phys. Rev. Lett.* **124**, 020503 (2020).
- [25] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, Recurrent neural network wave functions, *Phys. Rev. Res.* **2**, 023358 (2020).
- [26] C. Roth, A. Szabó, and A. H. MacDonald, High-accuracy variational Monte Carlo for frustrated magnets with deep neural networks, *Phys. Rev. B* **108**, 054410 (2023).
- [27] X. Liang, M. Li, Q. Xiao, J. Chen, C. Yang, H. An, and L. He, Deep learning representations for quantum many-body systems on heterogeneous hardware, *Mach. Learn.: Sci. Technol.* **4**, 015035 (2023).
- [28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2017), Vol. 30.
- [29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16×16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations* (2021).
- [30] I. von Glehn, J. S. Spencer, and D. Pfau, A self-attention ansatz for *Ab-initio* quantum chemistry, in *International Conference on Learning Representations* (2023). [OpenReview / ICLR Proceedings, Virtual Conference.].

- [31] G. Pescia, J. Nys, J. Kim, A. Lovato, and G. Carleo, Message-passing neural quantum states for the homogeneous electron gas, *Phys. Rev. B* **110**, 035108 (2024).
- [32] H. Shang, C. Guo, Y. Wu, Z. Li, and J. Yang, Solving Schrödinger equation with a language model, [arXiv:2307.09343](https://arxiv.org/abs/2307.09343).
- [33] L. L. Viteritti, R. Rende, and F. Becca, Transformer variational wave functions for frustrated quantum spin systems, *Phys. Rev. Lett.* **130**, 236401 (2023).
- [34] Y.-H. Zhang and M. Di Ventura, Transformer quantum state: A multipurpose model for quantum many-body problems, *Phys. Rev. B* **107**, 075147 (2023).
- [35] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, CvT: Introducing convolutions to vision transformers, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, Piscataway, NJ, 2021), pp. 22–31.
- [36] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, Cmt: Convolutional neural networks meet vision transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2022), pp. 12175–12185.
- [37] X. Cao, Z. Zhong, and Y. Lu, Vision transformer neural quantum states for impurity models, *Phys. Rev. B* **112**, 235155 (2025).
- [38] R. Rende and L. Loris Viteritti, Are queries and keys always relevant? A case study on transformer wave functions, *Mach. Learn.: Sci. Technol.* **6**, 010501 (2025).
- [39] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, Rethinking and improving relative position encoding for vision transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, Piscataway, NJ, 2021), pp. 10033–10041.
- [40] R. Rende, L. L. Viteritti, L. Bardone, F. Becca, and S. Goldt, A simple linear algebra identity to optimize large-scale neural network quantum states, *Commun. Phys.* **7**, 260 (2024).
- [41] L. L. Viteritti, R. Rende, A. Parola, S. Goldt, and F. Becca, Transformer wave function for the Shastry-Sutherland model: Emergence of a spin-liquid phase, *Phys. Rev. B* **111**, 134411 (2025).
- [42] S. Roca-Jerat, M. Gallego, F. Luis, J. Carrete, and D. Zueco, Transformer wave function for quantum long-range models, *Phys. Rev. B* **110**, 205147 (2024).
- [43] V. Herráiz-López, S. Roca-Jerat, M. Gallego, R. Ferrández, J. Carrete, D. Zueco, and J. Román-Roche, First- and second-order quantum phase transitions in the long-range unfrustrated antiferromagnetic Ising chain, *Phys. Rev. B* **111**, 014425 (2025).
- [44] K. Sprague and S. Czischek, Variational Monte Carlo with large patched transformers, *Commun. Phys.* **7**, 90 (2024).
- [45] H. Lange, G. Bornet, G. Emperauger, C. Chen, T. Lahaye, S. Kienle, A. Browaeys, and A. Bohrdt, Transformer neural networks and quantum simulators: A hybrid approach for simulating strongly correlated systems, *Quantum* **9**, 1675 (2024).
- [46] D. Hendrycks and K. Gimpel, Gaussian error linear units (GELUs), [arXiv:1606.08415](https://arxiv.org/abs/1606.08415).
- [47] S. Sorella, Green function Monte Carlo with stochastic reconfiguration, *Phys. Rev. Lett.* **80**, 4558 (1998).
- [48] L. L. Viteritti, R. Rende, and F. Becca (private communication).
- [49] D. Wu, R. Rossi, F. Vicentini, N. Astrakhantsev, F. Becca, X. Cao, J. Carrasquilla, F. Ferrari, A. Georges, M. Hibat-Allah *et al.*, Variational benchmarks for quantum many-body problems, *Science* **386**, 296 (2024).
- [50] A. Sinibaldi, D. Hendry, F. Vicentini, and G. Carleo, Time-dependent neural Galerkin method for quantum dynamics, *Phys. Rev. Lett.* **136**, 120402 (2026).
- [51] L. Gravina, V. Savona, and F. Vicentini, Neural projected quantum dynamics: A systematic study, *Quantum* **9**, 1803 (2024).
- [52] H. W. J. Blöte and Y. Deng, Cluster Monte Carlo simulation of the transverse Ising model, *Phys. Rev. E* **66**, 066110 (2002).
- [53] Y. Gu, W. Li, H. Lin, B. Zhan, R. Li, Y. Huang, D. He, Y. Wu, T. Xiang, M. Qin, L. Wang, and D. Lv, Solving the Hubbard model with neural quantum states, [arXiv:2507.02644](https://arxiv.org/abs/2507.02644).
- [54] P. Weinberg and M. Bukov, QuSpin: A Python package for dynamics and exact diagonalisation of quantum many body systems. Part I: Spin chains, *SciPost Phys.* **2**, 003 (2017).
- [55] A. Chen and C. Roth, Quantax: Flexible neural quantum states based on QuSpin, JAX, and Equinox, 2025, <https://github.com/ChenAo-Phys/quantax>.
- [56] M. Schmitt and M. Reh, jVMC: Versatile and performant variational Monte Carlo leveraging automated differentiation and GPU acceleration, *SciPost Phys. Codebases* **2** (2022).
- [57] A. Chen, V. Naik, and M. Heyl, Convolutional transformer wave functions [Data set], Zenodo, 2024, <https://doi.org/10.5281/zenodo.14035975>.