

Pareto Frontier of Neural Quantum States: Scalable, Affordable, and Accurate Convolutional Backflow for Strongly Correlated Lattice Fermions

Yuntian Gu ^{1,2,*} Zeyao Han ^{2,3,*} Wenrui Li^{1,2,*} Zhiyu Xiao ⁴ Tao Xiang ^{4,†} Mingpu Qin ^{5,‡}
Liwei Wang ^{1,§} and Dingshun Lv^{2,||}

¹State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology,
Peking University, Beijing 100871, China

²ByteDance Seed, Beijing, China

³Institute for Advanced Study, Tsinghua University, Beijing 100084, China

⁴Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China

⁵Key Laboratory of Artificial Structures and Quantum Control (Ministry of Education), School of Physics and Astronomy,
Shanghai Jiao Tong University, Shanghai 200240, China



(Received 10 February 2026; accepted 11 August 2026; published 6 October 2026)

The recent development of Neural Quantum States (NQS) has established them as one of the most accurate methods for studying strongly correlated many-fermion systems, outperforming existing many-body approaches for large systems. However, NQS calculations are currently highly resource intensive. In this work, we introduce an improved Pareto frontier of efficiency and accuracy for NQS in the simulation of strongly correlated lattice fermions. This frontier is defined by two complementary backflow-related architectures: the Sparse Convolutional Ansatz for Lattice Electrons (SCALE), which achieves high computational efficiency, and the Accurate Convolutional ansatz for lattice Electrons (ACE), which delivers high variational accuracy, based on benchmark results on the Hubbard model and the t - J model on large lattice sizes. SCALE utilizes a tailored convolutional design to enable an efficient local update based on the low-rank update of determinants. This structure reduces the computational scaling from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^3)$ in backflow methods, resulting in a practical speedup exceeding $40\times$ in our test cases while retaining accurate variational accuracy. As an application, we study the $1/8$ -doped pure Hubbard model up to size 32×32 and find no significant energy difference between the horizontal and vertical unit-filled stripe states, in contrast to the half-filled stripe state when next-nearest-neighbor hoppings are included. ACE, conversely, employs a deep convolutional stack to maximize expressive power, achieving high accuracy on large systems. We perform extensive benchmark calculations on the Hubbard and t - J models with the two architectures. On the challenging repulsive Hubbard model, SCALE achieves variational energies competitive with leading methods at a fraction of the computational cost. Meanwhile, ACE achieves lower variational energies than previously reported results while requiring only one-sixth of the run-time for system size 16×4 . These NQS approaches provide *scalable*, *affordable*, and *accurate* tools for exploring the rich physics of strongly correlated fermionic systems.

DOI: [10.1103/2kq2-5g5q](https://doi.org/10.1103/2kq2-5g5q)

I. INTRODUCTION

Understanding the emergence of exotic physics in strongly correlated fermion systems is a central and persistent challenge in condensed matter physics. Except for rare cases [1], exact analytical solutions are unattainable, necessitating the application of computational methods. However, strongly correlated systems usually harbor competing states with nearly

degenerate low energies, a feature that poses significant challenges for numerical simulations [2–4]. Consequently, results can be highly sensitive to finite-size effects and boundary conditions [5–9] as well as algorithmic biases, which may artificially favor one state over another. Over the past decades, a diverse array of numerical many-body methods has been developed, and benchmark calculations have been performed on iconic systems [10]. These methods collectively populate a Pareto front of computational tools, and the primary challenge remains to advance this frontier toward higher accuracy and greater efficiency.

Among existing numerical many-body methods, variational approaches are particularly powerful for determining the ground-state properties of strongly correlated quantum many-body systems. For instance, the density matrix renormalization group (DMRG) [11] method can provide nearly exact results for quasi-one-dimensional systems. However, the application of DMRG to two-dimensional (2D) systems suffers from the limited entanglement entropy encoded in

*These authors contributed equally to this work.

†Contact author: txiang@iphy.ac.cn

‡Contact author: qinmingpu@sjtu.edu.cn

§Contact author: wanglw@pku.edu.cn

||Contact author: ywlds163@gmail.com

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

the underlying wavefunction ansatz, i.e., the Matrix Product States (MPS) [12–14]. Although two-dimensional extension of MPS like projected entangled pair states (PEPS) [15] can encode the entanglement area law of the ground state of two-dimensional systems, they are often hindered by high computational complexity [14,16,17]. More recently, neural quantum states (NQS) have emerged as a highly flexible and expressive class of variational ansatz [18].

To effectively incorporate neural networks, approaches that directly generalize mean-field theories, such as Neural Network Backflow (NNBF) [19] and Hidden Fermion states [20], have proven to be highly accurate and efficient. Building upon these foundations, various architectures have been developed to better capture complex many-body correlations. For example, transformer backflow utilizes an attention mechanism to effectively capture long-range correlations, surpassing the best tensor network methods for large systems [21]. Conversely, the tensor-backflow method [22] generates the backflow Slater determinant based on tensor representation. Other specialized architectures, such as the Hidden Fermion Pfaffian State (HFPS) [23,24], offer fresh perspectives on complex phases by suggesting that superconductivity is enhanced by the next-nearest-neighbor hopping t' . Furthermore, recent studies have demonstrated that even simpler MultiLayer Perceptron (MLP) network backbones can achieve significantly enhanced accuracy when augmented with rigorous symmetry projections [25].

Despite these advancements, accurate NQS calculations remain extremely resource intensive. Developing more efficient and accurate NQS architectures is crucial for enabling their widespread application in the study of strongly correlated many-electron systems. In this work, that define an improved Pareto frontier for NQS simulations: the Sparse Convolutional Ansatz for Lattice Electrons (SCALE) and the Accurate Convolutional ansatz for lattice Electrons (ACE).

SCALE is engineered for high efficiency while remaining competitive in accuracy. It employs a convolutional layer that processes only local regions of the input at each time. This locality is the key to a highly optimized local update scheme, which leverages cached intermediate computations and fast low-rank matrix updates. This design reduces the computational scaling of the simulation from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^3)$, unlocking a practical speedup that exceeds $40\times$ for large systems. Complementing the efficient SCALE architecture, we introduce ACE to push the limits of variational accuracy on this frontier. ACE replaces SCALE's efficient layer with a deep stack of convolutional blocks, maximizing expressive power. While this design forfeits the fast update capability, it is built to further increase the accuracy over SCALE.

We perform extensive benchmark calculations on several canonical strongly correlated models. For the attractive Hubbard model, our results agree well with the intrinsically unbiased Determinant Quantum Monte Carlo (DQMC) simulation [26–29] and correctly identify the expected long-range s -wave pairing. On the more challenging repulsive Hubbard model, our framework demonstrates its utility. The efficient SCALE ansatz achieves a lower variational energy than several leading methods, including PEPS [30], tensor backflow [22], and HFPS [23]. It is also directly competitive with the top-tier transformer backflow [21], but at a fraction of the

computational cost. Meanwhile, the ACE results establish a new accuracy benchmark, surpassing recent results [25], while requiring only one-sixth of the computational cost. Crucially, the efficiency of SCALE allows us to address an open question concerning the orientation of stripe state that was computationally prohibitive for previous methods [21]; our simulations on 16×32 systems show that, in contrast to the partially filled stripe in the Hubbard model with next-nearest-neighboring hoppings [21], we find no significant difference in the energies for the horizontal and vertical unit-filled stripes in the pure Hubbard model with only nearest-neighboring hoppings. We further apply our method to the t - J model, obtaining energies comparable to large-scale DMRG on narrow cylinders [7] and lower than the DMRG results for wide cylinders. These comprehensive results establish SCALE and ACE as robust and efficient methods for scalable and high-accuracy simulations of strongly correlated systems.

II. METHODS

A. Variational Monte Carlo

The Variational Monte Carlo (VMC) method approximates the ground state of a quantum system by optimizing a parametrized trial wavefunction or ansatz, $|\psi_\theta\rangle$. According to the variational principle, the energy expectation value $E(\theta) = \langle \psi_\theta | \hat{H} | \psi_\theta \rangle / \langle \psi_\theta | \psi_\theta \rangle$ provides a rigorous upper bound to the true ground-state energy. In practice, this expectation value is computed stochastically by sampling configurations in the Fock basis $\{\mathbf{n}\}$ with respect to the probability distribution $P(\mathbf{n}) \propto |\psi_\theta(\mathbf{n})|^2$, where $\psi_\theta(\mathbf{n}) \equiv \langle \mathbf{n} | \psi_\theta \rangle$. This approach reformulates the energy expectation value as an average of the local energy, defined as $E_{\text{loc}} = \sum_{\mathbf{n}'} \psi_\theta(\mathbf{n})^{-1} \langle \mathbf{n} | \hat{H} | \mathbf{n}' \rangle \psi_\theta(\mathbf{n}')$, and can be obtained from configurations sampled using Markov chain Monte Carlo algorithms. The wavefunction parameters θ are then optimized using gradient-based methods to minimize the estimated energy, yielding the best variational approximation of the ground state. The details of the optimization method will be discussed in the following sections.

B. Sparse convolutional ansatz for lattice electrons

In this section, we introduce an NQS architecture that combines high expressive power with locality. A key requirement for efficient VMC simulations is the rapid calculation of wavefunction amplitude ratios, $\psi_\theta(\mathbf{n}')/\psi_\theta(\mathbf{n})$, for configurations $\mathbf{n}$ and $\mathbf{n}'$ that differ only locally. This operation is fundamental not only to the MCMC sampling process but also to the evaluation of the local energy for Hamiltonians containing local operators, such as hopping terms in the Hubbard model. Our architecture, most directly inspired by FiRE [31], is designed explicitly to take advantage of this locality.

The wavefunction ansatz takes the form of $\psi_\theta(\mathbf{n}) = \det[\Phi(\mathbf{n})]$, where the Slater matrix Φ is constructed from quasiparticle orbitals. These orbitals are not static but dynamically generated via a deep neural network that implements a backflow transformation [32–34], allowing the orbitals to depend on the particle configuration, and thereby capture complex electronic correlations. The generation of these configuration-dependent orbitals proceeds through a sequence of computational steps, as shown in Fig. 1(a).

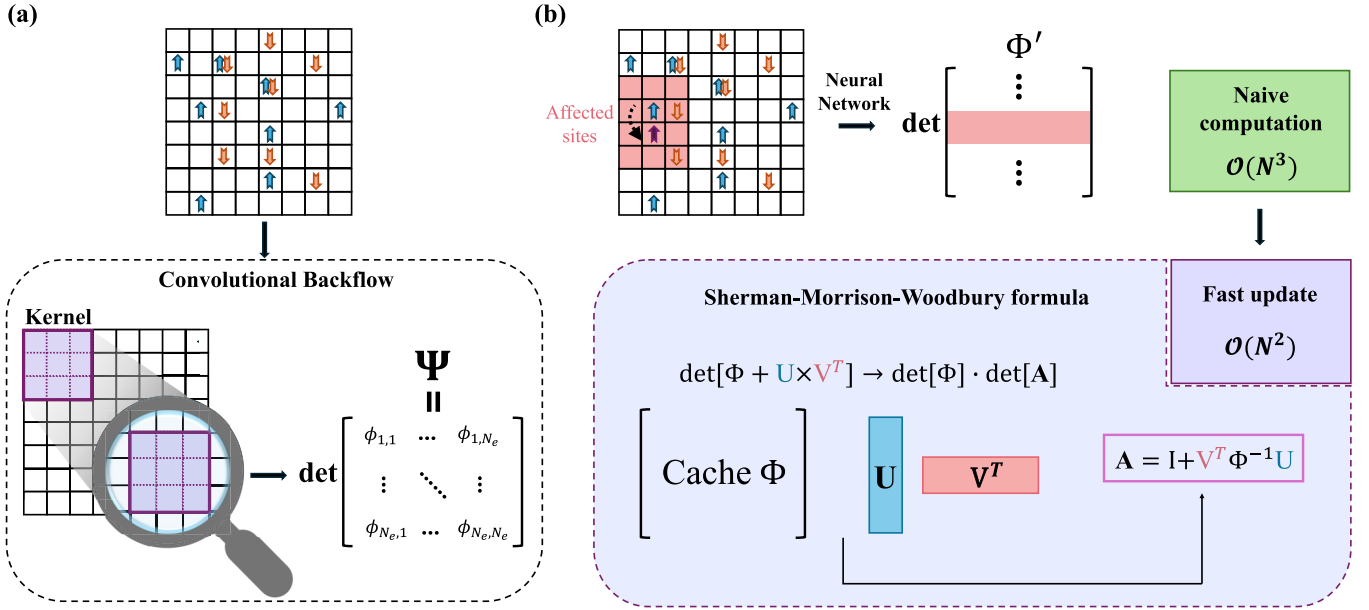


FIG. 1. Network architecture and efficient local updates. (a) Given an input configuration, a convolutional layer gathers information from local neighborhoods. The wavefunction is expressed as the determinant of the resulting backflow matrix. (b) SCALE enables efficient updates by only recomputing the orbitals of the affected sites. The computational cost of the determinant is reduced from $O(N^3)$ to $O(N^2)$ via the Sherman-Morrison-Woodbury formula.

(1) *Embedding layer*: Each lattice site $i \in \{1, \dots, L_x \times L_y\}$ has a discrete physical state determined by its occupancy with spin degrees of freedom. For the Hubbard model, the local Hilbert space is four dimensional, corresponding to the states $|0\rangle$, $|\uparrow\rangle$, $|\downarrow\rangle$, and $|\uparrow\downarrow\rangle$. We first map these physical states to high-dimensional vectors using a learnable embedding matrix $\mathbf{E}_i \in \mathbb{R}^{4 \times h}$, where h is the hidden dimension. Therefore, for a configuration $\mathbf{n}$, this gives an embedding matrix $\mathbf{H}^{(0)} = [\mathbf{E}_1(\mathbf{n}_1), \dots, \mathbf{E}_{L_x \times L_y}(\mathbf{n}_{L_x \times L_y})]^T \in \mathbb{R}^{(L_x \times L_y) \times h}$.

(2) *Convolutional layer*: The initial features are refined via a convolutional operation with residual connection [35]:

$$\mathbf{H}^{(1)} = \mathbf{H}^{(0)} + \sigma(\text{Conv}_{\theta_c}(\mathbf{H}^{(0)})). \quad (1)$$

Here, Conv_{θ_c} denotes a convolutional layer parametrized by $\theta_c = \{\mathbf{K}, b\}$ (convolution kernel $\mathbf{K} \in \mathbb{R}^{k \times k \times h \times h}$, with kernel size k , bias $b \in \mathbb{R}^h$), operating with the same boundary conditions as the Hamiltonian. The $k \times k$ kernel specifies the local neighborhood, and σ denotes the Sigmoid Linear Unit (SiLU) activation function [36].

(3) *Sitewise MLPs*: Each site's feature vector undergoes transformation through an MLP consisting of L_{MLP} fully connected layers, introducing nonlinear expressivity through a sequence of affine transformations and activations. For layer indices $l = 1, \dots, L_{\text{MLP}}$, the transformation follows:

$$\mathbf{H}^{(l+1)} = \sigma(\mathbf{H}^{(l)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}), \quad (2)$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{h \times h}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^h$ represent the weight matrix and the bias vector for the l th layer, respectively.

(4) *Orbital projection*: The final output $\mathbf{H}^{(L_{\text{MLP}}+1)}$ is transformed by a linear layer to produce backflow orbitals $M \in \mathbb{R}^{2N \times N_e}$, where $N = L_x \times L_y$, 2 denotes spin-up and spin-down, and N_e is the number of electrons. The Slater matrix Φ is generated by selecting the rows of M corresponding to the occupied sites.

This architectural sequence ensures that orbitals are dynamically generated according to the particle configuration, capturing nonlocal correlations while maintaining the locality required for efficient wavefunction update.

C. Efficient local updates

The design of our ansatz enables an optimized local update scheme, leveraging both cached intermediate results and low-rank matrix update techniques as shown in Fig. 1(b). In this section, we show the details of the mathematical formalism for these highly efficient implementations, which reduce the computational overhead of wavefunction updates during local energy evaluation and MCMC sampling.

(1) *Cached computation for neural network propagations*: To avoid redundant calculations when the particle configuration undergoes local changes, we maintain a cache of intermediate network outputs. For a configuration update from $\mathbf{n}$ to $\mathbf{n}'$, the cache allows us to recompute only the affected components of the orbital generation pipeline. We first identify sites where occupancy changes:

$$\text{exchange} = \{i \mid \mathbf{n}'_i \neq \mathbf{n}_i\}. \quad (3)$$

Due to the locality of convolutional layers, changes propagate only within a finite neighborhood determined by the size of the convolution kernel. For a convolutional kernel of size $k \times k$, the set of sites affected by changes in exchange is

$$\text{affect sites} = \bigcup_{i \in \text{exchange}} \mathcal{N}_k(i), \quad (4)$$

where $\mathcal{N}_k(i)$ denotes the neighborhood of site i within Chebyshev distance $(k-1)/2$. Instead of reevaluating the entire network, we only update features for affected sites. This

caching strategy reduces the computational complexity of network updates from $\mathcal{O}(N)$ to $\mathcal{O}(|\text{affect site}|)$.

(2) *Fast determinant updates via Sherman-Morrison-Woodbury formula:* For Slater matrix $\Phi(\mathbf{n}) \in \mathbb{R}^{N_e \times N_e}$, a local configuration change induces a low-rank modification [37]. We exploit this structure to update the determinants along with their inverses efficiently.

Let $\Phi' = \Phi(\mathbf{n}')$ denote the updated Slater matrix. The difference between Φ' and $\Phi (= \Phi(\mathbf{n}))$ lies in the rows corresponding to the changed indices:

$$\Phi' = \Phi + \mathbf{U}\mathbf{V}^\top, \quad (5)$$

where $\mathbf{U} \in \mathbb{R}^{N_e \times r}$ and $\mathbf{V} \in \mathbb{R}^{N_e \times r}$, with $r \ll N_e$. Using the Sherman-Morrison-Woodbury formula, the ratio of determinants and the inverse of the updated matrix are

$$\begin{aligned} \frac{\det(\Phi')}{\det(\Phi)} &= \det(\mathbf{A}), \\ (\Phi')^{-1} &= \Phi^{-1} - \Phi^{-1}\mathbf{U}\mathbf{A}^{-1}\mathbf{V}^\top\Phi^{-1}, \end{aligned} \quad (6)$$

where $\mathbf{A} = \mathbf{I}_r + \mathbf{V}^\top\Phi^{-1}\mathbf{U}$. Critically, both the determinant and the inverse of $\mathbf{A}$ can be derived from a single LU decomposition, further optimizing the calculation.

This combined strategy of cached network computations and low-rank matrix updates reduces the per-update complexity significantly, enabling large-scale VMC simulations.

D. Accurate convolutional ansatz for lattice electrons

We also introduce a second, computationally intensive ansatz: the ACE, which is designed to achieve high accuracy and establish definitive benchmarks. This architecture prioritizes maximal expressive power through increased network depth and a large neighborhood size, albeit at the cost of the efficient local update capabilities described in the previous section.

While ACE retains the overall structure of SCALE, it replaces the single convolutional layer [Eq. (1)] with a deep stack of L_{CNN} convolutional blocks. Starting from the initial embedding $\mathbf{H}^{(0)}$, the propagation through this deep stack proceeds as follows for each layer $l = 0, \dots, L_{\text{CNN}} - 1$:

$$\mathbf{H}^{(l+1)} = \text{LN}(\mathbf{H}^{(l)} + \sigma(\text{Conv}_{\theta_c}^{(l)}(\mathbf{H}^{(l)}))), \quad (7)$$

where $\text{Conv}_{\theta_c}^{(l)}$ denotes the l th convolutional layer and σ is the SiLU activation. We employ Layer Normalization (LN) [38] to stabilize the training across this deeper network. The final output $\mathbf{H}^{(L_{\text{CNN}})}$ is subsequently passed to the sitewise MLPs and orbital projection defined in Sec. II B.

The primary trade-off for the increased accuracy in ACE is a reduction in computational efficiency. As the number of layers L_{CNN} increases, the network's effective range expands significantly. In a sufficiently deep network, a local change at a single site i propagates through the layers, altering the feature vectors $\mathbf{H}^{(L_{\text{CNN}})}$ across all lattice sites. This global dependency destroys the locality property required for the cached computation scheme [Eq. (4)]. Consequently, any update from configuration $\mathbf{n}$ to $\mathbf{n}'$ necessitates a full reevaluation of the network pipeline. Thus, ACE serves as a powerful, high-accuracy architecture that can take advantage of the trade-off between expressivity and efficiency.

E. Comparison to existing architectures

To contextualize the advancements of SCALE and ACE, it is instructive to compare their fundamental design principles with other recent competitive approaches, such as transformer backflow [21] and Hidden Fermion methods like HFPS [23]. The Hidden Fermion approach is conceptually completely distinct from the NNBF paradigm that underpins SCALE and ACE. Specifically, HF methods construct wavefunctions by evaluating Slater determinants in an augmented Hilbert space involving “hidden” additional fermionic degrees of freedom. These states are then projected back onto the physical Hilbert space through a constraint that is optimized using a neural network parametrization. In contrast, backflow-based methods operate strictly within the original physical Hilbert space, using neural networks to dynamically generate configuration-dependent, multiparticle orbitals.

Within the backflow paradigm, the defining feature of SCALE is its resolution of the asymptotic scaling bottlenecks inherent to NQS. In standard NNBF architectures, the neural network introduces global dependencies: A local change in a single electron's position instantly alters the hidden features of all spatial orbitals. This inflicts a severe computational penalty, necessitating a full $\mathcal{O}(N)$ reevaluation of the neural network for every single local update. By explicitly restricting the network's receptive field using a tailored convolutional geometry, SCALE ensures that physical updates only perturb a mathematically bounded local neighborhood. This architectural sparsity allows for exact caching, slashing the network forward-pass cost from $\mathcal{O}(N)$ down to an $\mathcal{O}(1)$ operation per update.

On the other hand, ACE adopts a design choice that departs from prevailing machine learning practice. While global attention mechanisms (such as vision transformers [39]) have become the default for 2D structured data, they effectively discard the strict spatial geometry of the lattice in favor of fully connected, all-to-all graphs. Reverting to deep convolutional stacks instead of attention is a highly nontrivial choice in modern machine learning, but it is uniquely suited to the problem domain in this work. Strongly correlated lattice fermions are governed by strictly local operators (e.g., nearest-neighbor hopping t and on-site repulsion U), meaning long-range correlations build up spatially through local correlations. ACE respects this physical inductive bias by employing a deep stack of localized convolutional blocks. This forces the network to extract features hierarchically, matching the physical reality of the lattice, rather than relying on naive all-to-all attention. In this sense, ACE resembles the spirit of renormalization. ACE demonstrates the fundamental physical insight that deep, geometrically constrained convolutions are actually more effective at propagating information and capturing ground-state correlations for these specific local Hamiltonians than global attention mechanisms. By doing so, ACE achieves high variational accuracy, outperforming both transformer backflow and HFPS, while maintaining a lighter computational footprint than full-attention models.

F. Optimization and ground-state projection

We use the pretrain method [21] to provide NQS with better initialization. The variational parameters of the

TABLE I. Comparison of ground-state energy for the 16×16 Hubbard model. Our SCALE and SCALE + GFMC are among the most accurate methods, while ACE and ACE + GFMC methods consistently find the lowest variational and projected energies, giving the lowest energies among the methods compared for this system. For results where statistical errors are not shown, they are one or several orders of magnitude smaller than the last displayed significant digit. The lowest variational energy and the lowest projected energy in each column are shown in bold.

Systems	$t' = 0$, OBC	$t' = 0$, PBC	$t' = -0.2$, PBC
PEPS [30]	-0.7260(2)		
HFPS + sym [24]		-0.7515(1)	-0.7360(1)
Tensor backflow [22]	-0.7219	-0.7509	-0.7335
Tensor backflow + Lanczos [22]	-0.7257	-0.7552	-0.7386
Transformer [21]	-0.7275	-0.7563	
SCALE	-0.7236	-0.7529	-0.7373
SCALE + GFMC	-0.7267	-0.7560	-0.7405
ACE	-0.7280	-0.7573	-0.7430
ACE + GFMC	-0.7288	-0.7583	-0.7440

wavefunction are then optimized within the VMC framework using the MARCH optimizer, a sophisticated stochastic gradient method designed for the stable and efficient training of NQS [21]. To further improve upon the variational result, we subsequently employ Green's function Monte Carlo (GFMC), a projector method that filters out residual excited-state contributions from the optimized VMC wavefunction under the fixed-node approximation [40,41]. Throughout this paper, we have ensured sufficiently large Monte Carlo sampling such that, unless explicitly denoted with parentheses, the statistical error bars are one or several orders of magnitude smaller than the precision of the reported numbers in all tables and smaller than the data markers in all figures.

III. RESULTS

To illustrate the performance of SCALE and ACE, we perform extensive calculations for the two-dimensional Hubbard model [42] and for a reduced t - J model; the benchmark results are summarized in Tables I–III. The Hamiltonians con-

TABLE II. Energy of horizontal vs vertical stripes on the 32×16 Hubbard model with PBC. The two systems shown are a unit-filled stripe at $t' = 0$ and a half-filled stripe at $t' = -0.2$. The results from both SCALE and SCALE + GFMC show a clear energy preference for the horizontal stripe in the half-filled case, while there is no significant difference between the horizontal and vertical stripe energies within our accuracy for the unit-filled stripe case with $t' = 0$. Statistical errors in the energies are one or several orders of magnitude smaller than the last displayed significant digit.

Systems	$t' = 0$, unit-filled stripe ($\lambda = 8$)	$t' = -0.2$, half-filled stripe ($\lambda = 4$)
SCALE, horizontal	-0.7507	-0.7371
SCALE, vertical	-0.7507	-0.7359
SCALE + GFMC, horizontal	-0.7546	-0.7406
SCALE + GFMC, vertical	-0.7547	-0.7397

sidered here are the Hubbard model

$$H_h = -t \sum_{\langle i,j \rangle, \sigma} (c_{i\sigma}^\dagger c_{j\sigma} + \text{H.c.}) - t' \sum_{\langle\langle i,j \rangle\rangle, \sigma} (c_{i\sigma}^\dagger c_{j\sigma} + \text{H.c.}) + U \sum_i n_{i\uparrow} n_{i\downarrow} \quad (8)$$

and the reduced t - J model

$$H_{t-J} = -t \sum_{\langle i,j \rangle, \sigma} \mathcal{P}_G (c_{i\sigma}^\dagger c_{j\sigma} + \text{H.c.}) \mathcal{P}_G - t' \sum_{\langle\langle i,j \rangle\rangle, \sigma} \mathcal{P}_G (c_{i\sigma}^\dagger c_{j\sigma} + \text{H.c.}) \mathcal{P}_G + J \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j, \quad (9)$$

respectively. In these Hamiltonians, $\langle ij \rangle$ denotes nearest-neighbor sites and $\langle\langle ij \rangle\rangle$ denotes next-nearest-neighbor sites. $c_{i\sigma} (c_{i\sigma}^\dagger)$ is the annihilation (creation) operator for spin species σ at site i . $\mathbf{S}_i$ is the spin operator at site i . We set the nearest-neighbor hopping amplitude $t = 1$ as the unit of energy and primarily focus on the $1/8$ hole doping. In the reduced t - J Hamiltonian, $\mathcal{P}_G = \prod_i (1 - n_{i\uparrow} n_{i\downarrow})$ is the Gutzwiller projection operator to remove double occupancy. Notice that the t - J model from large U expansion of Hubbard model also includes a density-density term $-\frac{J}{4} \sum_{\langle i,j \rangle} n_i n_j$ and other terms.

A. Computational complexity

We first analyze the computational scaling of NQS by identifying the primary bottleneck: MCMC sampling and local energy evaluation, which requires $\mathcal{O}(N)$ steps of local update. In global architectures like ACE or transformers, any local change to the configuration propagates through the entire network. This necessitates a full reevaluation of the wavefunction for every proposed update. Since the determinant calculation scales as $\mathcal{O}(N^3)$, performing $\mathcal{O}(N)$ updates yields an asymptotic theoretical complexity of $\mathcal{O}(N^4)$ per sweep. However, in practice, the total computational cost must account for the neural network overhead. For ACE, the total cost takes the form of $aN^2 + bN^4$, where the aN^2 term arises from the neural network evaluations [e.g., convolutional neural network

TABLE III. The comparison of energies for the t - J model with $J/t = 1/3$, $t'/t = 0.2$, $h_m = 0.5$ at $1/8$ doping. For wider cylinders, SCALE and ACE achieve comparable and lower energies to DMRG with bond dimension $D = 30\,000$ [7]. Statistical errors in the energies are one or several orders of magnitude smaller than the last displayed significant digit. The lowest variational energy and the lowest projected energy in each column are shown in bold.

Systems	32×4	32×6	16×8	16×10
DMRG	-0.5071	-0.4994	-0.5000	-0.4961
SCALE	-0.5025	-0.4926	-0.4974	-0.4974
SCALE + GFMC	-0.5044	-0.4942	-0.4986	-0.4985
ACE	-0.5066	-0.4964	-0.5001	-0.5001
ACE + GFMC	-0.5067	-0.4967	-0.5003	-0.5003

(CNN) processing] across N updates, and the bN^4 term originates from the determinant calculations. Because the neural network prefactor a is relatively large, the asymptotic $\mathcal{O}(N^4)$ limit is not yet fully dominant for the finite system sizes studied here. This interplay results in the observed intermediate empirical scaling of roughly $\mathcal{O}(N^{2.19})$.

In contrast, SCALE is explicitly designed to overcome this bottleneck. By leveraging cached convolutional computations, the feature update cost becomes $\mathcal{O}(1)$. The locality of this ansatz ensures that the determinant can be updated iteratively using the Sherman-Morrison-Woodbury formula, reducing the cost from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ for each update [37]. Consequently, the total theoretical complexity for an optimization step of $\mathcal{O}(N)$ updates is reduced to $\mathcal{O}(N^3)$. Analogous to the global case, the practical cost for SCALE is a combination of efficient local neural network updates and the iterative determinant updates, yielding a total cost model of $aN + bN^3$. Within our tested system sizes, the efficient $\mathcal{O}(N)$ feature updates heavily mitigate the overall operational cost, resulting in an effective empirical scaling of approximately $\mathcal{O}(N^{1.65})$.

Figure 2 validates this analysis with numerical data, showing the scaling of cost for each optimization step for SCALE,

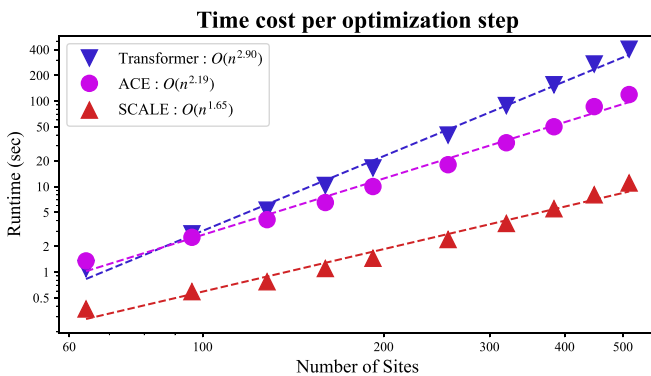


FIG. 2. Time cost per optimization step as a function of system size. The plot compares the run-time scaling of SCALE against global architectures (ACE and transformer) under a fixed batch size. SCALE achieves a highly favorable asymptotic complexity [empirically $\approx \mathcal{O}(N^{1.65})$], resulting in substantial speedups for larger systems compared to the steeper scaling of ACE [$\approx \mathcal{O}(N^{2.19})$] and transformer [$\approx \mathcal{O}(N^{2.90})$].

ACE, and transformer backflow [21]. As we can see from the figure, for the largest system size tested ($N = 512$), SCALE completes an optimization step in approximately 11 s. In comparison, the ACE architecture requires about 120 s, while the transformer needs over 400 s, resulting in speedups of roughly $10\times$ and $40\times$, respectively. The run-time measurements demonstrate that SCALE maintains a significantly lower effective scaling exponent ($N^{1.65}$) compared to the ACE ($N^{2.19}$) and transformer ($N^{2.90}$) baselines, indicating that the efficiency gain becomes increasingly dramatic for large-scale simulations.

B. Attractive Hubbard model

To show the accuracy of our NQS architectures, we first apply them to the attractive Hubbard model with negative U . This model is a canonical benchmark for s -wave superconductivity, where the attractive on-site interaction induces the formation of Cooper pairs [43,44].

A standard method for quantifying this state in a particle-conserving system is through the pair-pair correlation function, $C_p(x, 0) = \langle \hat{\Delta}^\dagger(x) \hat{\Delta}(0) \rangle$, where $\hat{\Delta}(x) = c_{x\downarrow} c_{x\uparrow}$. The long-range behavior of $C_p(x, 0)$ as $x \rightarrow \infty$ serves as a direct proxy for the off-diagonal long-range order [45] that characterizes the superconducting phase. Crucially, the attractive Hubbard model is free of the fermion sign problem in the spin-balanced sector, which allows DQMC simulations to provide numerically exact, unbiased results. The energies from ACE and SCALE agree perfectly well with the QMC results, and the comparison can be found in Table IV.

In Fig. 3, we plot pair correlations $C_p(x, 0)$ as a function of distance x for two different system sizes and interaction strengths. We study both the weak-coupling regime ($U = -2$) and the strong-coupling regime ($U = -8$), choosing $t' = 0$ and $1/8$ hole doping. The results demonstrate outstanding accuracy for both ACE and SCALE. Across all systems and parameters tested, the correlations calculated by both SCALE and ACE are almost indistinguishable from the numerically exact DQMC benchmarks. The perfect alignment confirms that our methods can faithfully capture the complex quantum correlations of the s -wave superconducting ground state.

The high accuracy of SCALE and ACE on the negative U model is supported by an analysis detailed in Appendix A. We show that, through a particle-hole transformation (PHT), a BCS wavefunction can be explicitly mapped to our wave-

TABLE IV. Comparison of the energies in the attractive Hubbard model. The energies from SCALE and ACE agree perfectly well with DQMC. For results where statistical errors are not shown, they are one or several orders of magnitude smaller than the last displayed significant digit.

Systems	$8 \times 8, U = -2$	$8 \times 8, U = -8$	$12 \times 12, U = -2$	$12 \times 12, U = -8$
DQMC	-2.0343	-4.017(1)	-2.0416	-4.016(1)
SCALE	-2.0337	-4.0161	-2.0404	-4.0141
SCALE + GFMC	-2.0338	-4.0167	-2.0406	-4.0158
ACE	-2.0338	-4.0165	-2.0405	-4.0159
ACE + GFMC	-2.0339	-4.0168	-2.0407	-4.0161

function ansatz. This confirms that the generalized backflow transformation provides the necessary flexibility to accurately describe the superconducting ground state [34] without requiring explicit pairing terms.

C. Repulsive Hubbard model

We now turn to the 2D repulsive Hubbard model, an iconic model for strongly correlated many-body physics [42] and is believed to be relevant for Cuprate superconductors [46]. We focus on the most challenging and intriguing regime with $U = 8$, and hole doping $\delta = 1/8$. We compare the ACE and SCALE results with several leading methods to demonstrate their superior accuracy and efficiency.

1. Accuracy on 16×4 systems

We focus first on a relatively small system with size 16×4 under periodic boundary conditions (PBC). In Fig. 4(a), we plot the ground-state energy against the total computational cost, comparing the energies from ACE with previous state-of-the-art results. As it is well established that explicitly projecting a wavefunction onto the correct symmetry subspace can significantly lower the variational energy [23,25,47], we also evaluate the energy in the fully symmetric sector to provide a fair comparison against other symmetrized methods.

Our standard ACE architecture, trained without explicit symmetries, achieves a ground-state energy of $-0.76464(1)$.

This result is already more accurate than that of HFPS [23] while requiring only one-fifteenth of the cost. By directly optimizing the projected, symmetrized state, we obtain a variational energy of $-0.76623(1)$. This surpasses the previous best result $-0.766073(6)$ from NNBF [25] with only one-sixth of the computational cost. Additionally, a GFMC calculation with the ACE trial wavefunction gives an even lower energy of $-0.76647(2)$. Details of the symmetry projection process can be found in Appendix B.

2. Accuracy on large systems

We then study a large system with size 32×8 and including the next-nearest-neighboring hopping $t' = -0.2$, a challenging regime where the trade-off between variational accuracy and computational cost is critical. Our new results with SCALE and ACE and the comparison with the transformer backflow results [21] are summarized in Fig. 4(b), where the ground-state energy against computational cost is plotted. For the sake of simplicity, from here we report energies without symmetry projection. We want to emphasize that the energies can be further lowered by performing symmetry projections.

We further push the system size to as large as 32×32 (with $t' = 0$), achieving a variational ground-state energy of -0.7501 , and can be further reduced to -0.7548 using GFMC. The unit-filled stripe pattern for this repulsive Hub-

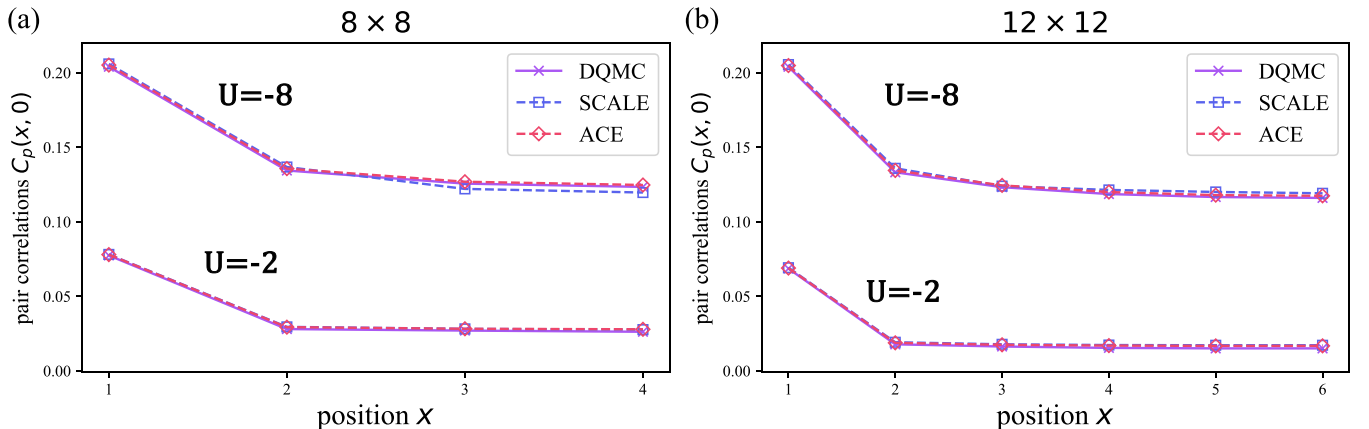


FIG. 3. Pair correlations $C_p(x, 0)$ for the attractive Hubbard model. The plots show results for (a) 8×8 and (b) 12×12 systems at weak ($U = -2$) and strong ($U = -8$) interaction strengths, with $t' = 0$. The results from our ACE and SCALE models perfectly align with the exact DQMC benchmarks, demonstrating their high accuracy. Statistical uncertainties from Monte Carlo sampling are smaller than the marker size in all panels.

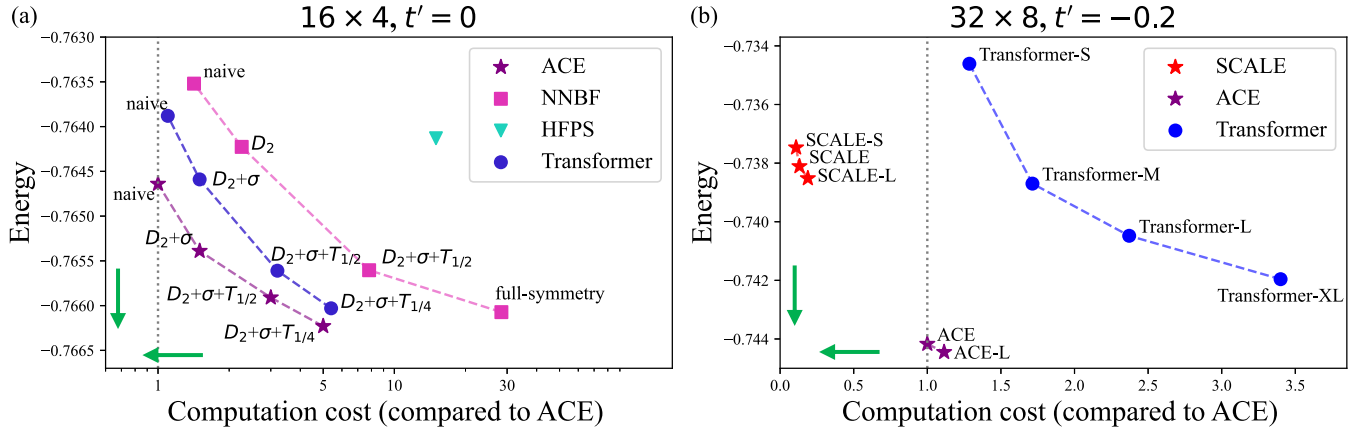


FIG. 4. Accuracy vs cost trade-off on the $U = 8$ Hubbard model. The plots compare variational ground-state energy against relative computational cost. (a) For the 16×4 ($t' = 0$) system, we analyze the effect of progressively enforcing symmetries on the ACE architecture (purple stars). Labels denote the cumulative addition of point group (D_2), spin-flip (σ), half-translation ($T_{1/2}$), and quarter-translation ($T_{1/4}$) symmetries. ACE achieves lower energies than transformer [21], HFPS [23], and NNBF [25] while maintaining a fraction of the computational cost of the full-symmetry NNBF. (b) On the larger 32×8 ($t' = -0.2$) system, the ACE and SCALE variants establish an improved Pareto front, outperforming transformer-based models [21] in both accuracy and computational efficiency.

bard system is shown in Appendix D. This confirms the scalability of our method.

As Fig. 4(b) shows, our ACE and SCALE results clearly establish an improved Pareto front, outperforming the previous state-of-the-art transformer-based architectures [21] by a large margin. The plot highlights two distinct advantages offered by our architectures. SCALE is highly efficient, being $10\times$ faster than the small transformer while achieving a better variational energy. ACE achieves a substantially lower ground-state energy than any other method. Remarkably, it reaches this new level of accuracy while also being more computationally efficient than all the tested transformer architectures. These results confirm that our proposed new architectures offer a significantly better trade-off between cost and accuracy. SCALE enables rapid computations with high efficiency, whereas ACE establishes state-of-the-art accuracy without sacrificing practical run-time. Details of the optimizer settings and of the configurations for each architecture are listed in Tables V and VI, respectively.

3. Comparison with other leading methods

In this subsection, we show a comparison of our results with other leading methods for a large-scale 16×16 system. We consider systems under different boundary conditions,

i.e., open boundary conditions (OBC) and PBC, and systems with and without t' . As shown in Table I, both SCALE and ACE demonstrate exceptional performance. SCALE is highly competitive, delivering accuracy on par with previous leading methods, including PEPS [30], HFPS [24], tensor backflow [22], and transformer [21]. In the meantime, ACE gives the lowest variational energies among the methods compared here, confirming the capabilities of our method in large-scale simulation of strongly correlated quantum many-body systems.

4. Orientation of stripe order

In a previous work [21], it was found that the half-filled stripe state for the Hubbard model with $t' = -0.2$ tends to align along the horizontal direction when the studied system has a rectangular geometry (i.e., the length L_x is greater than the width L_y). In this context, a horizontal stripe refers to states where the rivers of holes run along the longer x direction, whereas in a vertical stripe, the river of holes runs along the shorter y direction. The stripe filling denotes the number of doped holes per unit length along this one-dimensional river of holes; a unit-filled stripe has one doped hole per lattice spacing along the stripe. Reference [21] reported that the vertical stripe has a significantly higher energy. Their analysis

TABLE V. Hyperparameters for the MARCH optimizer. Shared parameters apply to both models, while model-specific settings are listed in their respective columns.

Hyperparameter	ACE	SCALE
Learning rate schedule (t)	$10^{-1}(1 + \max(t - 1000, 0)/4000)^{-1}$	$5 \times 10^{-1}(1 + \max(t - 1000, 0)/8000)^{-1}$
Optimizer momentum (μ, β)		0.95, 0.995
Damping (λ)		0.001
Batch size		$\text{clip}(32L_x \times L_y, 4096, 8192)$
Local energy clipping		5.0
MCMC steps per sample		$2.5L_x \times L_y$
Total optimization steps		50 000

TABLE VI. Model hyperparameters for SCALE and ACE variants. We list the details of the convolutional hidden dimension (h_{conv}), MLP hidden dimension (h_{MLP}), number of convolutional layers (L_{conv}), and number of MLP layers (L_{MLP}). Note that SCALE-L uses a convolutional filter with a neighborhood range of Euclidean distance $L_2 \leq 2$, as opposed to the $L_2 \leq \sqrt{2}$ range used in other models. The SCALE-XL hyperparameters are used for 32×32 system size only.

Model	h_{conv}	h_{MLP}	L_{conv}	L_{MLP}	Kernel Shape
SCALE-S	256	512	1	4	
SCALE	512	512	1	4	
SCALE-L	384	512	1	4	
SCALE-XL	640	1280	1	4	
ACE	144	256	16	2	
ACE-L	192	256	16	2	

showed that a horizontal stripe allows the doped holes to delocalize over a longer contiguous distance, which provides a more favorable gain in kinetic energy, while the potential energy difference between the two orientations remains small. It was proposed [21] to also study the preference of stripe orientation of the unit-filled stripe in the pure Hubbard model without t' in finite systems. However, such a comparison requires simulating much larger systems (e.g., a 32×16 lattice for $1/8$ doping), which was computationally intractable using the transformer backflow with the resources available then. The efficiency of our SCALE architecture makes these large-scale simulations feasible. We calculate the energies for both the horizontal and vertical unit-filled stripe states for the 32×16 system, and the results are listed in Table II. The temporary pinning-field procedure used to initialize the desired stripe orientation is described in Appendix C. For the $t' = -0.2$ case, we find similar results; i.e., the horizontal stripe has a lower energy. However, for the $t' = 0$ case, the energies for the horizontal and vertical stripe states are comparable within the accuracy of our method. A possible reason for this result is that the wavelength for the unit-filled stripe at $1/8$ is already very large ($\lambda = 16$ for spin density), so the kinetic energies of the doped hole in a hole river with length 16 and 32 have little difference.

D. t - J model

We also apply our methods to the t - J model defined in Eq. (9), where the density-density term $-\frac{J}{4} \sum_{\langle i,j \rangle} n_i n_j$ is not included.

The crucial feature of the t - J model is the constraint of no double occupancy, which is enforced by the Gutzwiller projection operator, $\mathcal{P}_G = \prod_i (1 - n_{i\uparrow} n_{i\downarrow})$, as shown in Eq. (9). This operator projects the wavefunction onto the subspace of states without double occupancy. In our variational Monte Carlo framework, this constraint is implemented explicitly. The physical wavefunction amplitude is computed as $\Psi(\mathbf{n}) = \mathcal{P}_G \psi_\theta(\mathbf{n})$. In practice, this means that any proposed configuration $\mathbf{n}$ containing a doubly occupied site is immediately discarded by setting its amplitude to zero.

We follow the setup in a recent large-scale DMRG calculation in Ref. [7] and set the parameters to $J/t = 1/3$,

$t'/t = 0.2$, and a hole-doping level of $\delta = 1/8$, to study a situation that is mathematically equivalent to electron-doped cuprates ($t'/t = -0.2$, $1/8$ electron doping) via a particle-hole transformation in the corresponding Hubbard model. We cannot directly study electron-doped carriers in this t - J Hilbert space because double occupancy is excluded. The simulations are performed on cylindrical boundary conditions. To help stabilize the expected order, an antiferromagnetic pinning field with strength $h_m = 0.5$ is applied to the first column of sites on the left open edge. Table III presents our results, comparing SCALE and ACE (with and without GFMC projection) to highly accurate DMRG data (with bond dimension $D = 30000$) for 32×4 , 32×6 , 16×8 , and 16×10 lattices. As expected, DMRG is more accurate in narrow cylinder systems with widths 4 and 6. However, when the width of the system is increased to 8, the SCALE and ACE results are comparable to the DMRG energy. For an even wider cylinder with width 10, the SCALE and ACE energies are much lower than the DMRG energy.

It was found that $1/8$ doping is likely to lie near the boundary between the Neel antiferromagnetic phase at lower doping and a unit-filled stripe phase at higher doping levels in the t - t' - J model [7]. In Fig. 5, we compare the energy, spin, and hole density of the 16×10 t - t' - J model calculated with ACE for both the unit-filled stripe state and the Neel antiferromagnetic state. We find that the unit-filled stripe state has a lower energy. Within the reduced model considered here, our results support placing this parameter point near or on the stripe side of the transition as electron doping is increased. We leave the determination of the precise phase boundary in this model for future investigations.

These results show that our architectures are highly accurate for the t - J model, especially for larger (wider) system sizes. It was found in Refs. [5–9] that large finite-size effects exist in the Hubbard and t - J model when the t' term is included. The SCALE and ACE methods therefore provide alternative accurate tools for determining the ground state of these models in the thermodynamic limit.

IV. CONCLUSION AND PERSPECTIVES

In this work, we have introduced a powerful and complementary pair of neural quantum state architectures, SCALE and ACE, which together establish a new Pareto frontier of the NQS for the simulation of strongly correlated many-fermion systems on lattice. This framework directly addresses the fundamental trade-off between computational efficiency and variational accuracy that has long defined the limits of the field. By providing two specialized tools, one for scalability (SCALE) and one for high accuracy (ACE), our work offers a flexible and robust toolkit to address key challenges in condensed matter physics.

SCALE provides an algorithmic improvement. Its tailored convolutional design, combined with the fast low-rank update scheme, successfully breaks the $\mathcal{O}(N^4)$ scaling barrier of conventional backflow methods, achieving a much more favorable $\mathcal{O}(N^3)$ complexity. This asymptotic improvement, which translates to a practical speedup exceeding $40\times$ for 32×16 system, is not merely an incremental gain. It fun-

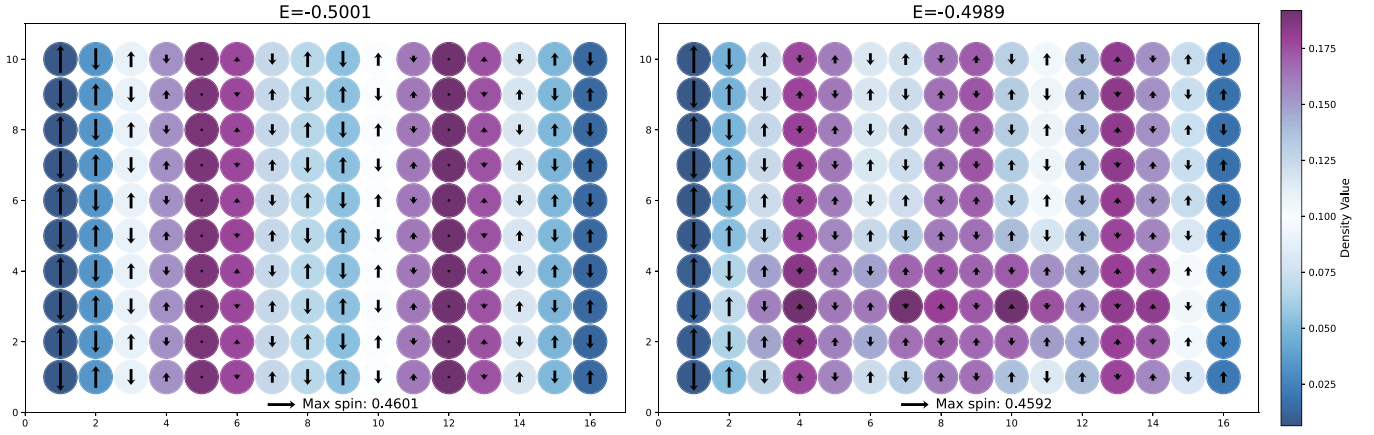


FIG. 5. Comparison between the stripe phase and the Neel state for the 16×10 t - J model. The parameters are set to $J/t = 1/3$, $t'/t = 0.2$ with $1/8$ hole doping (corresponding to the $J/t = 1/3$, $t'/t = -0.2$, $1/8$ electron-doped case), and $h_m = 0.5$ under cylindrical boundary conditions. (Left) The true ground state exhibits a clear unit-filled stripe pattern with a lower energy ($E = -0.5001$). (Right) A competing Neel phase lies higher in energy ($E = -0.4989$). For both plots, the circle color represents the local hole density, while the arrows indicate the magnitude and direction of the local spin density.

damentally alters the landscape of what is computationally feasible. We demonstrated this directly by addressing an open question regarding the orientation of stripe order of the pure Hubbard model (with only nearest-neighboring hopping t) on a 32×16 lattice, a system size previously intractable for other NQS methods with high accuracy. Our findings suggest that, unlike partially filled stripes in the Hubbard model with next nearest-neighboring hoppings [21], the horizontal and vertical unit-filled stripes in the pure Hubbard model have comparable energy within the accuracy of our method. Moreover, with the favorable scaling of SCALE, we can reach the large 32×32 system for the pure Hubbard model.

Furthermore, the high accuracy of SCALE, transformer backflow, challenges the prevailing assumption that mechanisms capturing long-range interactions at the input layer are a prerequisite for highly accurate ground-state simulations. Our results suggest that a deep yet local convolutional backflow is exceptionally effective at propagating information and capturing the complex, long-range correlation characteristic of the ground state for strongly correlated systems.

On the other end of the frontier, ACE provides high variational accuracy. By replacing the efficient sparse layer with a deep convolutional stack, ACE maximizes expressive power. As shown in Fig. 4 and Table I, it consistently gives lower variational energies than the methods compared here, including transformer-based models [21] and enhanced backflows [25]. Crucially, ACE achieves this new level of accuracy while being more computationally efficient than its competitors, requiring, for example, only one-sixth of the run-time of NNBF for a more accurate result for a system with size 16×4 . This combination of precision and relative efficiency is vital for the study of strongly correlated systems, where the central challenge, as noted in our introduction, is often the resolution of competing states with nearly degenerate energies [3]. ACE provides a powerful and reliable new tool for this exact purpose.

In summary, we have introduced a unified (SCALE/ACE) framework that advances the state of the art in variational simulations of strongly correlated lattice fermions, provid-

ing a scalable path forward. The efficiency of SCALE, with its $\mathcal{O}(N^3)$ scaling, enables systematic, large-scale studies of finite-size effects, allowing for more reliable extrapolations to the thermodynamic limit. Simultaneously, the accuracy of ACE makes it an ideal instrument for a definitive search for elusive phases, such as d -wave superconductivity, in the repulsive Hubbard and t - J models. Future work can leverage this platform to explore key physical problems, such as the microscopic mechanism of high- T_c cuprate superconductors. Extending these methods beyond ground-state calculations presents another promising research direction.

ACKNOWLEDGMENTS

We thank ByteDance Seed for inspiration and encouragement, and Hang Li for his guidance and support. We thank Yuan-Yao He and Hao Du for providing DQMC data. We thank Yang Shen for providing DMRG data. L.W. was supported by the National Science and Technology Major Project (2022ZD0114902) and the National Natural Science Foundation of China (Grants No. 92470123 and No. 62276005). M.Q. acknowledges the support from the National Key Research and Development Program of MOST of China (2022YFA1405400), the National Natural Science Foundation of China (Grants No. 12274290 and No. 12522406), and the Innovation Program for Quantum Science and Technology (2021ZD0301902).

DATA AVAILABILITY

The NQS calculations use the LaQX codebase [48]. The data supporting this work are openly available in Ref. [49].

APPENDIX A: REPRESENTABILITY OF PAIRING WAVEFUNCTIONS

A general mean-field wavefunction can be written as the Thouless state, $|\psi\rangle = \exp(\frac{1}{2} \sum_{p,q} F_{pq} \hat{c}_p^\dagger \hat{c}_q^\dagger) |0\rangle$, where p, q

runs over all orbital and spin indices and F is an antisymmetric matrix. The BCS state, in particular, can be viewed as a special instance. The ground state of the attractive Hubbard model, despite its many-body nature, is believed to be close to such mean-field wavefunctions.

When benchmarking our method on the attractive Hubbard model, we use a PHT defined by $U = \prod_{i=1}^N (c_{i\downarrow}^\dagger + (-1)^i c_{i\downarrow})$, where N is the number of sites and $(-1)^i$ gives opposite signs to different sublattices. Under this PHT, we have $c_{i\uparrow} \rightarrow c_{i\uparrow}$, $c_{i\downarrow} \rightarrow (-1)^i c_{i\downarrow}^\dagger$, and the Hubbard Hamiltonian transforms as $\hat{H} \rightarrow \hat{H}_{\text{PH}} = -t \sum_{\langle i,j \rangle, \sigma} (c_{i\sigma}^\dagger c_{j\sigma} + \text{H.c.}) + U \sum_i n_{i\uparrow} (1 - n_{i\downarrow})$.

Specifically, a BCS wavefunction $|\psi_{\text{BCS}}\rangle \propto \exp(\sum_{i,j} F_{ij} \hat{c}_{i\uparrow}^\dagger \hat{c}_{j\downarrow}^\dagger) |0\rangle$ is transformed to a particle-number-conserved one

$$|\Psi_{\text{BCS}}\rangle \propto \exp\left(\sum_{i,j=1}^N (-1)^j F_{i,j} c_{i\uparrow}^\dagger c_{j\downarrow}^\dagger\right) \prod_{l=1}^N c_{l\downarrow}^\dagger |0\rangle. \quad (\text{A1})$$

This is a Slater determinant, as can be verified using the identity

$$\begin{aligned} & \exp\left(\sum_{i,j=1}^N M_{i,j} c_{i\uparrow}^\dagger c_{j\downarrow}^\dagger\right) c_{l\downarrow}^\dagger \exp\left(-\sum_{i,j=1}^N M_{i,j} c_{i\uparrow}^\dagger c_{j\downarrow}^\dagger\right) \\ &= c_{l\downarrow}^\dagger + \sum_i c_{i\uparrow}^\dagger M_{il}. \end{aligned} \quad (\text{A2})$$

By taking $M_{ij} = (-1)^j F_{i,j}$, we have

$$\begin{aligned} \langle \mathbf{n} | \Psi_{\text{BCS}} \rangle &\propto \langle \mathbf{n} | \prod_{l=1}^N \left(c_{l\downarrow}^\dagger + \sum_i c_{i\uparrow}^\dagger M_{il} \right) | 0 \rangle \\ &= \det \left[\mathbf{n} \star \begin{pmatrix} M \\ I \end{pmatrix} \right], \end{aligned} \quad (\text{A3})$$

where $\mathbf{n} \star$ denotes selecting N rows of the subsequent matrix based on the occupied orbitals of $\mathbf{n}$. This form is well within the representability of our NQS ansatz, which motivates us to simulate the particle-hole transformed Hamiltonian $\hat{H}_{\text{PH}}$ instead, with the intuition that the s -wave pairing presented in this model is better captured.

APPENDIX B: SYMMETRY PROJECTION

The symmetrized wavefunction, ψ_{sym} , is obtained by projecting the initial wavefunction:

$$\psi_{\text{sym}} = \sum_g \Pi(\mathbf{n}, g) \chi_g \psi(g\mathbf{n}), \quad (\text{B1})$$

where $g \in G$ is an operation from the symmetry group, $\Pi(\mathbf{n}, g) = \pm 1$ is the fermionic sign from permutations, and χ_g is the character of the symmetry representation.

Our procedure involves two stages. In the first stage, we perform 3×10^4 initial steps without enforcing any symmetry. The resulting state is then projected onto the fully symmetric state, a step with essentially no computational cost. This baseline projection corresponds to “naive” in Fig. 4(a).

In the second stage, we gradually enforce symmetry following the procedure in Ref. [50] through a sequence of training steps, as shown in Fig. 6. First, we restore D_2 and

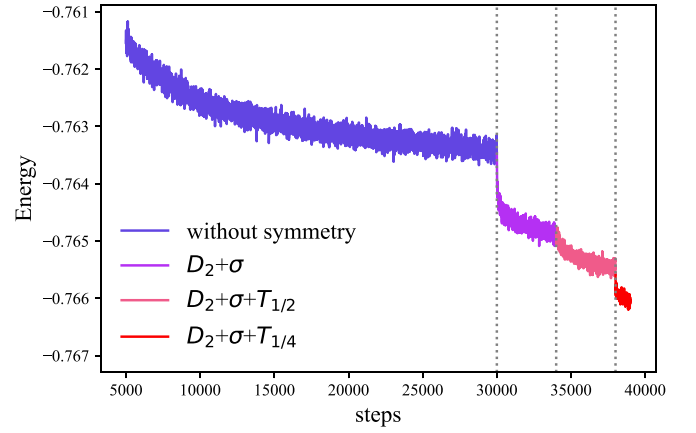


FIG. 6. Variational energy of ACE as a function of optimization steps. The system is a Hubbard model with size 16×4 and $t' = 0$. We can see that symmetry projection greatly increases the variational accuracy.

spin inversion symmetry and train for 4×10^3 steps. Next, we additionally restore the half-translation symmetries and train for another 4×10^3 steps. Finally, we restore the 1/4-translational symmetries and train for 10^3 steps. After this sequential training, the state is once again projected onto the fully symmetric state at essentially no cost.

APPENDIX C: PINNING FIELD

To avoid convergence to local minima during optimization, temporary pinning fields are applied following Ref. [21]. Different types of pinning fields are imposed to get desired patterns for a simple backflow network. Next, the pattern is transferred to SCALE or ACE using a pretraining method. The final optimization with MARCH [21] is performed with the pinning field removed, allowing the wavefunction to relax and find the true ground state of the original Hamiltonian. With this method, we can stabilize vertical or horizontal stripe states with different fillings or wavelengths.

APPENDIX D: ADDITIONAL DATA

We provide additional benchmark results for the attractive Hubbard model in Table IV. The table compares the ground-state energies from our models (SCALE, ACE, and their GFMC variants) against established DQMC benchmarks for various system sizes and interaction strengths. We also provide the ground-state charge and spin density for the repulsive ($U = 8$) 32×32 Hubbard model in Fig. 7.

APPENDIX E: HYPERPARAMETER

For full reproducibility, this Appendix outlines the specific parameters used in our experiments. We provide the settings for the MARCH optimizer, followed by the architectural details for the SCALE and ACE model variants.

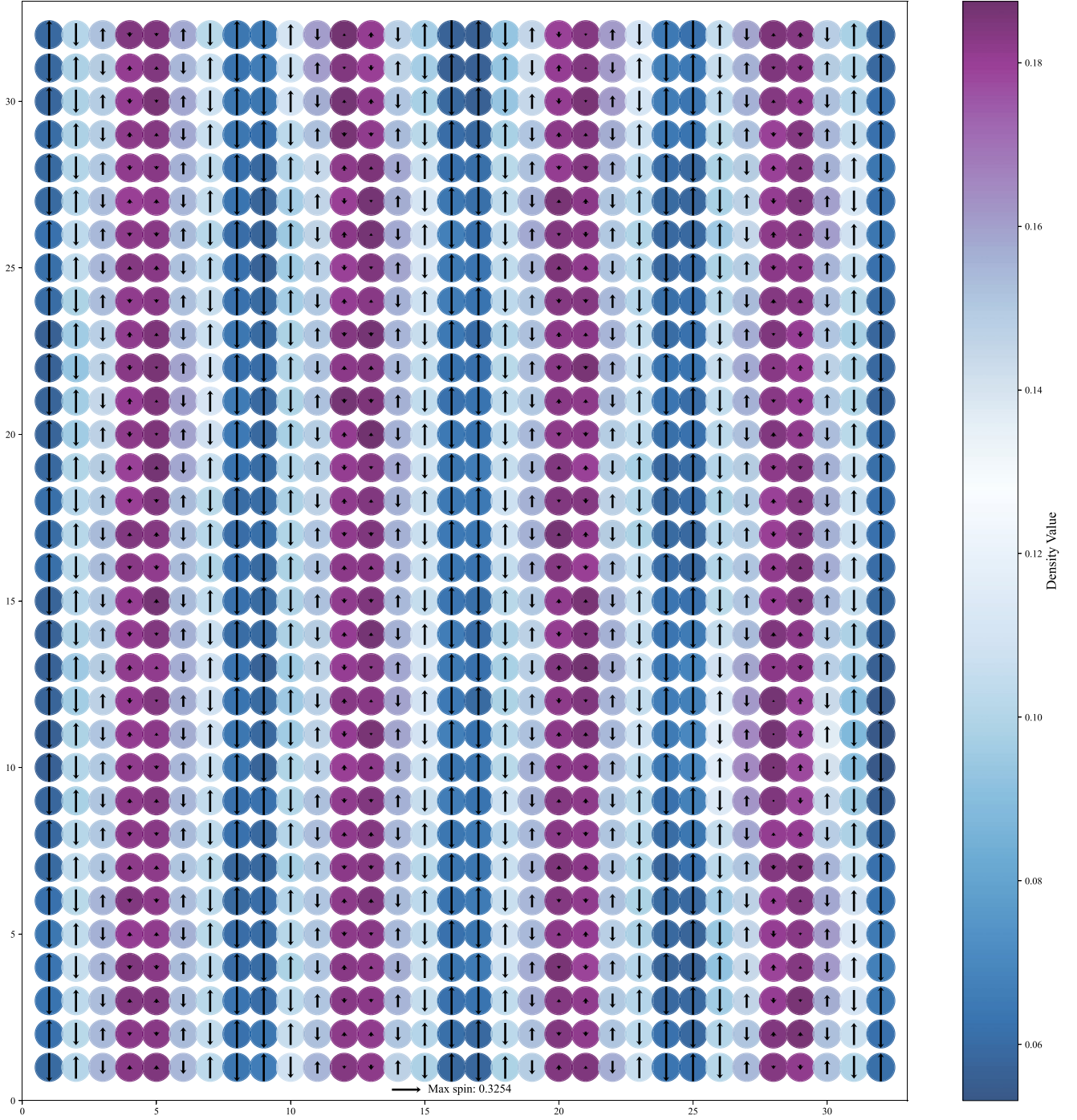


FIG. 7. Ground-state density and spin distribution for the repulsive ($U = 8$) 32×32 Hubbard model with $t' = 0$ at 1/8 doping under PBC. The color of the circles indicates the local hole density, and the arrows represent the magnitude and direction of the local spin density. We can see a clear unit-filled stripe state from the results.

-
- [1] E. H. Lieb and F. Y. Wu, Absence of Mott transition in an exact solution of the short-range, one-band model in one dimension, *Phys. Rev. Lett.* **20**, 1445 (1968).
 - [2] E. Fradkin, S. A. Kivelson, and J. M. Tranquada, Colloquium: Theory of intertwined orders in high temperature superconductors, *Rev. Mod. Phys.* **87**, 457 (2015).
 - [3] M. Qin, T. Schäfer, S. Andergassen, P. Corboz, and E. Gull, The Hubbard model: A computational perspective, *Annu. Rev. Condens. Matter Phys.* **13**, 275 (2022).
 - [4] D. P. Arovas, E. Berg, S. A. Kivelson, and S. Raghu, The Hubbard model, *Annu. Rev. Condens. Matter Phys.* **13**, 239 (2022).

- [5] C.-M. Chung, M. Qin, S. Zhang, U. Schollwöck, and S. R. White (The Simons Collaboration on the Many-Electron Problem), Plaquette versus ordinary d -wave pairing in the t' -Hubbard model on a width-4 cylinder, *Phys. Rev. B* **102**, 041106 (2020).
- [6] H. Xu, C.-M. Chung, M. Qin, U. Schollwöck, S. R. White, and S. Zhang, Coexistence of superconductivity with partially filled stripes in the Hubbard model, *Science* **384**, eadh7691 (2024).
- [7] Y. Shen, X. Qian, and M. Qin, Ground state of electron-doped t - t' - j model on cylinders: An investigation of finite size and boundary condition effects, *Chin. Phys. B* **34**, 087105 (2025).
- [8] X. Lu, F. Chen, W. Zhu, D. N. Sheng, and S.-S. Gong, Emergent superconductivity and competing charge orders in hole-doped square-lattice t - j model, *Phys. Rev. Lett.* **132**, 066002 (2024).
- [9] H.-C. Jiang, T. P. Devereaux, and S. A. Kivelson, Competition between charge-density-wave and superconducting orders on eight-leg square Hubbard cylinders, *arXiv:2511.18644*.
- [10] J. P. LeBlanc, A. E. Antipov, F. Becca, I. W. Bulik, G. K.-L. Chan, C.-M. Chung, Y. Deng, M. Ferrero, T. M. Henderson, C. A. Jiménez-Hoyos, *et al.*, Solutions of the two-dimensional Hubbard model: Benchmarks and results from a wide range of numerical algorithms, *Phys. Rev. X* **5**, 041041 (2015).
- [11] S. R. White, Density matrix formulation for quantum renormalization groups, *Phys. Rev. Lett.* **69**, 2863 (1992).
- [12] S. Östlund and S. Rommer, Thermodynamic limit of density matrix renormalization, *Phys. Rev. Lett.* **75**, 3537 (1995).
- [13] U. Schollwöck, The density-matrix renormalization group in the age of matrix product states, *Ann. Phys.* **326**, 96 (2011).
- [14] J. I. Cirac, D. Pérez-García, N. Schuch, and F. Verstraete, Matrix product states and projected entangled pair states: Concepts, symmetries, theorems, *Rev. Mod. Phys.* **93**, 045003 (2021).
- [15] F. Verstraete and J. I. Cirac, Renormalization algorithms for quantum-many body systems in two and higher dimensions, *arXiv:cond-mat/0407066*.
- [16] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, *Ann. Phys.* **349**, 117 (2014).
- [17] T. Xiang, *Density Matrix and Tensor Network Renormalization* (Cambridge University Press, Cambridge, England, 2023).
- [18] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
- [19] D. Luo and B. K. Clark, Backflow transformations via neural networks for quantum many-body wave functions, *Phys. Rev. Lett.* **122**, 226401 (2019).
- [20] J. Robledo 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).
- [21] 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, *Nat. Commun.* **17**, 7838 (2026).
- [22] X. Liang, Investigating the Fermi-Hubbard model by the tensor-backflow method, *Phys. Rev. B* **113**, 155121 (2026).
- [23] A. Chen, Z.-Q. Wan, A. Sengupta, A. Georges, and C. Roth, Neural network-augmented Pfaffian wave-functions for scalable simulations of interacting fermions, *arXiv:2507.10705*.
- [24] C. Roth, A. Chen, A. Sengupta, and A. Georges, Superconductivity in the two-dimensional Hubbard model revealed by neural quantum states, *arXiv:2511.07566*.
- [25] K. Loehr and B. K. Clark, Enhancing neural network backflow, *arXiv:2510.26906*.
- [26] R. Blankenbecler, D. J. Scalapino, and R. L. Sugar, Monte Carlo calculations of coupled boson-fermion systems. I, *Phys. Rev. D* **24**, 2278 (1981).
- [27] J. E. Hirsch, Two-dimensional Hubbard model: Numerical simulation study, *Phys. Rev. B* **31**, 4403 (1985).
- [28] J. E. Hirsch, Discrete Hubbard-Stratonovich transformation for fermion lattice models, *Phys. Rev. B* **28**, 4059 (1983).
- [29] Y.-Y. He, H. Shi, and S. Zhang, Reaching the continuum limit in finite-temperature *ab initio* field-theory computations in many-fermion systems, *Phys. Rev. Lett.* **123**, 136402 (2019).
- [30] W.-Y. Liu, H. Zhai, R. Peng, Z.-C. Gu, and G. K.-L. Chan, Accurate simulation of the Hubbard model with finite fermionic projected entangled pair states, *Phys. Rev. Lett.* **134**, 256502 (2025).
- [31] M. Scherbela, N. Gao, P. Grohs, and S. Günnemann, Accurate *ab-initio* neural-network solutions to large-scale electronic structure problems, *arXiv:2504.06087*.
- [32] E. Wigner, On the interaction of electrons in metals, *Phys. Rev.* **46**, 1002 (1934).
- [33] R. P. Feynman and M. Cohen, Energy spectrum of the excitations in liquid helium, *Phys. Rev.* **102**, 1189 (1956).
- [34] L. F. Tocchio, F. Becca, A. Parola, and S. Sorella, Role of backflow correlations for the nonmagnetic phase of the t - t' Hubbard model, *Phys. Rev. B* **78**, 041101 (2008).
- [35] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 770–778.
- [36] S. Elfving, E. Uchibe, and K. Doya, Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, *Neural Netw.* **107**, 3 (2018).
- [37] S. R. White, D. J. Scalapino, R. L. Sugar, E. Y. Loh, J. E. Gubernatis, and R. T. Scalettar, Numerical study of the two-dimensional Hubbard model, *Phys. Rev. B* **40**, 506 (1989).
- [38] J. L. Ba, J. R. Kiros, and G. E. Hinton, Layer normalization, *arXiv:1607.06450*.
- [39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, An image is worth 16×16 words: Transformers for image recognition at scale, in *Proceedings of the International Conference on Learning Representations (ICLR)* (ICLR, 2021).
- [40] N. Trivedi and D. M. Ceperley, Ground-state correlations of quantum antiferromagnets: A Green-function Monte Carlo study, *Phys. Rev. B* **41**, 4552 (1990).
- [41] D. F. B. ten Haaf, H. J. M. van Bommel, J. M. J. van Leeuwen, W. van Saarloos, and D. M. Ceperley, Proof for an upper bound in fixed-node Monte Carlo for lattice fermions, *Phys. Rev. B* **51**, 13039 (1995).
- [42] J. Hubbard, Electron correlations in narrow energy bands, *Proc. R. Soc. London A* **276**, 238 (1963).
- [43] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Microscopic theory of superconductivity, *Phys. Rev.* **106**, 162 (1957).
- [44] R. T. Scalettar, E. Y. Loh, J. E. Gubernatis, A. Moreo, S. R. White, D. J. Scalapino, R. L. Sugar, and E. Dagotto, Phase

- diagram of the two-dimensional negative- u Hubbard model, *Phys. Rev. Lett.* **62**, 1407 (1989).
- [45] C. N. Yang, Concept of off-diagonal long-range order and the quantum phases of liquid He and of superconductors, *Rev. Mod. Phys.* **34**, 694 (1962).
- [46] F. Zhang and T. Rice, Effective Hamiltonian for the superconducting Cu oxides, *Phys. Rev. B* **37**, 3759 (1988).
- [47] L. Sharma, A. Shokry, R. Nutakki, O. Simard, M. Ferrero, and F. Vicentini, Comparing symmetrized determinant neural quantum states for the Hubbard model, *Phys. Rev. B* **113**, 245104 (2026).
- [48] ByteDance, LAQX, Github, 2026, <https://github.com/bytedance/laqx>.
- [49] Y. Gu, Data for “Pareto frontier of neural quantum states: Scalable, affordable, and accurate convolutional backflow for strongly correlated lattice fermions”, Github, 2026, https://github.com/guyuntian/open_data.
- [50] L. L. Viteritti, R. Rende, A. Parola, S. Goldt, and F. Becca, Transformer wave function for two dimensional frustrated magnets: Emergence of a spin-liquid phase in the Shastry-Sutherland model, *Phys. Rev. B* **111**, 134411 (2025).