


Unifying view of fermionic neural network quantum states: From neural network backflow to hidden fermion determinant states

Zejun Liu  and Bryan K. Clark*

The Anthony J. Leggett Institute for Condensed Matter Theory and IQUIST and NCSA Center for Artificial Intelligence Innovation and Department of Physics, University of Illinois at Urbana-Champaign, Illinois 61801, USA

 (Received 19 January 2024; revised 17 June 2024; accepted 13 August 2024; published 12 September 2024)

Among the variational wave functions for fermionic Hamiltonians, neural network backflow (NNBF) and hidden fermion determinant states (HFDS) are two prominent classes that provide accurate approximations to the ground state. Here we develop a unifying view of fermionic neural quantum states casting them all in the framework of NNBF. NNBF wave functions have configuration-dependent single-particle orbitals (SPO) which are parameterized by a neural network. We show that HFDS can be written as a NNBF wave function with a restricted low-rank r additive correction to the SPO times a neural-network generated determinant of an $r \times r$ matrix. Furthermore, we show that in NNBF wave functions, this $r \times r$ determinant can generically be removed when r is less than or equal to the number of fermions, at the cost of further complicating the additive SPO correction increasing its rank by r . We numerically and analytically compare additive SPO corrections generated by the product of two matrices with inner dimension r . We find that larger r wave functions span a larger space and give evidence that simpler and more direct updates to the SPO's tend to be more expressive and better energetically. These suggest the standard NNBF approach is preferred amongst other related choices. Finally, we uncover that the row selection used to select single-particle orbitals allows significant sign and amplitude modulation between nearby configurations and is partially responsible for the quality of NNBF and HFDS wave functions.

DOI: [10.1103/PhysRevB.110.115124](https://doi.org/10.1103/PhysRevB.110.115124)

I. INTRODUCTION

Simulating quantum many body systems is difficult often requiring various approximations to make progress. Since the early history of quantum mechanics, variational approaches have been one of these core approximations. In the variational approach, one starts with a parameterized class of wave functions and searches amongst this class for the lowest energy state. Hartree Fock, a variational search over the class of noninteracting wave functions was amongst the earliest variational approaches for simulating fermions, and works by optimizing the parameterized single particle orbitals (SPO) which make up a Slater determinant [1]. Even today, many of the state-of-the-art approaches for fermions and frustrated magnets build on top of this noninteracting class dressing and complementing it in various ways. For example, a standard class of Fermionic variational wave functions is the Slater-Jastrow form, $\Psi(c) = J(c)\Psi_{SD}(c)$, where $\Psi_{SD}(c)$ is a noninteracting Slater determinant and $J(c)$ is a multiplicative Jastrow factor which adds significant correlation on top of the Slater determinant [2–4]. The Jastrow factor $J(c)$ has historically been a strictly positive function taking the form of an exponential of one and two-body operators $J(c) = \exp[-u_1(c) + u_2(c)]$. More recently, with advances in optimization as well as the progress in machine learning (ML) architectures which approximate a wide class of functions

[5–18], Jastrows have become more sophisticated and parameterized, spanning wave function classes such as correlated product states, RBM, etc. [13,16–22].

In the context of Fermion systems, to further improve the expressibility of variational wave functions, recent work has applied ML architectures to dress the Slater determinant part of the Slater-Jastrow wave function. Two such paradigms include the neural network backflow (NNBF) [11–13,23–25] and the hidden Fermion determinant state (HFDS) [7,8]. NNBF replaces the static set of SPO with a configuration-dependent set which are generated by a neural network. HFDS instead works in the paradigm of projected hidden fermions using neural networks to replace the standard Slater determinant with a larger determinant which includes SPO's from an additional projected r hidden fermions. These seemingly distinct wave functions both achieve similar significant improvements to the variational power for fermionic systems. One cannot help but inquire about the existence of some underlying mechanism that links these states.

In this work, we investigate explicitly the connection between NNBF and HFDS. Surprisingly, when the number of hidden fermions is less than the number of electrons, $r \leq N_s$, HFDS can be written exactly in a NNBF form as a configuration dependent update to the SPO (see Fig. 1). This set of configuration dependent SPO's can either be written as an additive low-rank $2r$ correction to the static SPO's or alternatively as a SPO correction generated from first doing a rank r additive correction followed by a (right) multiplicative correction. The way these corrections are generated by

*Contact author: bkclark@illinois.edu

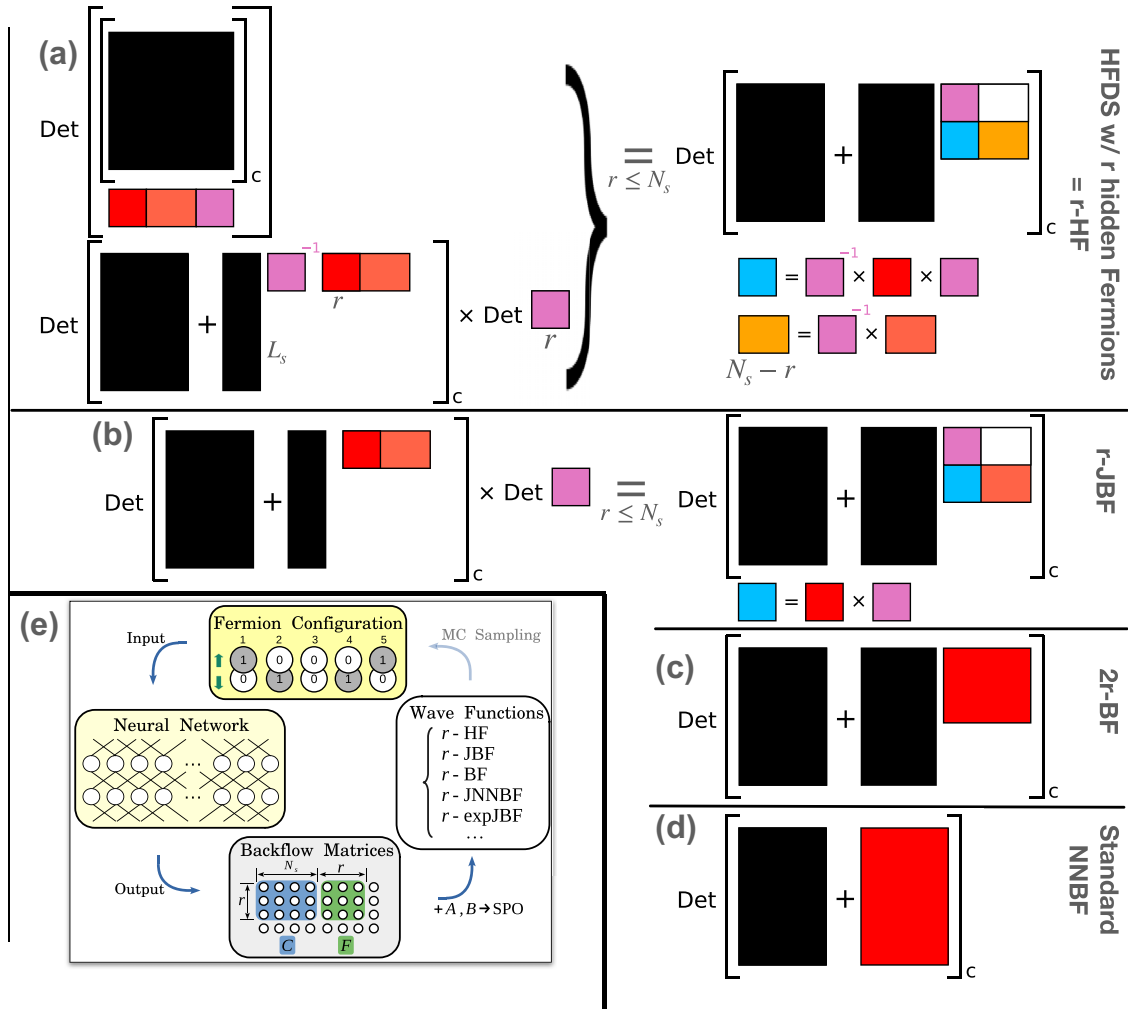


FIG. 1. [(a)–(d)] Visualization of wave functions. Red area are configuration-dependent matrices output from the NN; other colored area depend on those outputs; black area are optimized but configuration-independent and white area is zero. Same color in a given line are identical (up to a trivial added identity). Brackets with c in the lower-right indicate row selection of that matrix with configuration c . All wave functions (including HFDS) can be written in the style of NNBF as an additive correction to the SPO. Additive correction to lower wave functions are less constrained and this additional freedom is often reflected in better energetics. Note that (c) is formally a subset of (d) even when $r > L_s$. (b) is a subset of (c) when the NN for (c) has two additional layers and otherwise we see that both (b) and (c) and (a) and (b) can be optimized towards each other with low infidelity. The determinant factor in the left-hand side of (b) can be replaced by an exponential Jastrow with energetics of similar quality. (e) wave-function architecture for various NNBF: fermion configuration c as a set of binary numbers is input to a feed-forward neural network with two hidden layers. The output from the NN gives the $r \times N_s$ -size $C_\theta(c)$ and $r \times r$ -size $F_\theta(c)$ backflow matrices, which are then combined with the static matrices A and B to form SPO for various NNBF wave functions. The amplitudes from the wave functions are used to sample new fermion configurations with Monte Carlo.

HFDS involve using the NN output in a slightly complicated fashion (i.e., taking inverse of some matrices and multiplying them by others); we numerically demonstrate that these more complicated low-rank corrections are worse than simply doing a direct low-rank corrections and analytically show that direct low-rank corrections are always worse than the standard NNBF. This gives evidence that the type of NNBF induced by HFDS is less expressive than the standard NNBF when the number of hidden fermions is less than the physical fermion number.

When $r > N_s$, we can no longer view HFDS as just NNBF; instead, we have to view it as NNBF (still with an additive rank- r correction to the SPO's) times an NN-generated $r \times r$ determinant. This multiplicative factor acts like a Jastrow

factor albeit one that can be negative; toward that end, we call it a determinant factor. We find numerically that (fixing the NNBF SPO update), using either a determinant factor, a standard Jastrow factor or a standard Jastrow factor times an additional configuration-dependent sign given similar energetics. When we allow the NNBF SPO update to be the standard (i.e., full rank), we find little to no energetic improvement to including any multiplicative factor.

Finally, we investigate what feature gives NNBF (equivalently HFDS) its effective representability of low-energy ground states. Toward that end, we rewrite a number of wave functions as a standard (fixed, nonconfiguration dependent) Slater determinant times a NN-generated multiplicative factor. We find that the NN-generated multiplicative factor coming

from NNBF differs from other similar-looking wave functions by its rapid oscillations in both sign and magnitude, a feature that we attribute to the row selection of the determinant in these wave functions.

II. NEURAL NETWORK BACKFLOW

Both the NNBF and the HFDS build on top of Slater determinants. A Slater determinant is represented by a set of N_s single particle orbitals (where N_s is the total number of fermions), where each of them is scalar-valued function in space. On lattice systems, the SPO is defined by a vector on the lattice sites. In the continuum, it is defined by a fixed set of numbers which specify the three-dimensional (or two-dimensional) orbitals; this may be coefficients of a Fourier expansion or points on a tricubic spline. Our work will focus on the case of the lattice but the argument we make holds generically where in the continuum the neural network instead outputs the finite coefficients necessary to generate the three-dimensional orbitals.

On a discrete lattice the SPO's are a $L_s \times N_s$ matrix A , where $L_s = (2s + 1)L$ and L is the number of lattice sites or spatial orbitals and s is the spin of fermions. Let $A[c]$ be the $N_s \times N_s$ matrix generated by taking the rows of A which correspond to the locations of fermions in a configuration c (equivalently evaluating the SPO function with respect to the fermion coordinates in the continuum). The amplitude of a Slater determinant wave function for c is then $\Psi_{\text{SD}}(c) = \det A[c]$.

In NNBF, we instead make the SPO configuration-dependent,

$$\Psi(c) = J(c) \det(\tilde{A}_\theta(c)[c]), \quad (1)$$

where the subscript θ are parameters to the neural network (NN) which takes as input a configuration c and outputs a matrix (or a series of matrices) which is used to construct $\tilde{A}_\theta(c)$. We will see examples where $\tilde{A}_\theta(c)$ is built from a product or sum of other matrices and then the neural network outputs the configuration-dependent matrices from which the product or sum is taken. Take $J(c) = 1$ when it is not specified.

The NN needs to be permutation invariant with respect to the input; in standard NNBF, this network is a feed forward neural network (FFNN) which takes the input c as an L_s bit binary number with 1's at the sites the fermions occupy. To make comparisons fair, throughout this paper we fix all the layers (number of layers and layer widths w) except for a final linear layer whose output changes to be the full set of relevant matrices (in Sec. VB, we will discuss the possibility of reconstructing the SPO for one wave function with another by adding a few more layers, otherwise, we assume the NN's have identical architecture).

There are various ways the new SPO can be generated from the output of the NN. For example, through an additive correction

$$\tilde{A}_\theta(c) = A + M_\theta(c), \quad (2)$$

where $M_\theta(c)$ is an $L_s \times N_s$ matrix. Alternatively, a left rotation, $\tilde{A}_\theta(c) = L_\theta(c)A$ (shown in Ref. [11] to be nearly identical to an additive correction) or a right rotation $\tilde{A}_\theta(c) = A\tilde{F}_\theta(c)$ can be used to generate the new SPO. Left and right

rotations are qualitatively distinct with left rotations mixing sites within each individual single particle orbital and right rotations mixing SPO between each other at fixed site. We can even compose changes to the SPO doing first an additive correction and then a right rotation,

$$\tilde{\tilde{A}}_\theta(c) = (A + M_\theta(c))\tilde{F}_\theta(c). \quad (3)$$

Interestingly, right rotations can be factored out of the Slater determinant and treated as a determinant factor of the form

$$J_{\det}(c) = \det \tilde{F}_\theta(c). \quad (4)$$

The converse of this is also true: any determinant factor $\det F_\theta(c)$ for any $F_\theta(c)$ whose size $r \leq N_s$ can be absorbed into Eq. (3) by letting \tilde{F} be F padded with the identity up to size N_s . This will lead us to interchangeably treat a determinant factor and a final right rotation of the SPO as equivalent (paying attention when necessary that $r \leq N_s$) throughout. For example, the wave function of Eq. (3) will be JNNBF,

$$\Psi_{\text{TNNBF}} = \Psi_{\text{JNNBF}} \equiv J_{\det} \Psi_{\text{NNBF}}. \quad (5)$$

The NN parameters θ and A are all optimized when applied to the ground state approximation for Fermionic Hamiltonian. Later we will consider additional static matrices (e.g., B) which like A do not depend on the neural network but whose coefficients are optimized.

III. HIDDEN FERMION DETERMINANT STATES

In HFDS, starting with an $L_s \times N_s$ set of SPO's A , the HFDS supplements the physical system with the introduction of r hidden fermions and consequently an additional r single particle orbitals as well as some number of new sites. The r additional single particle orbitals on the original "real" sites are represented by a static $L_s \times r$ matrix B and $B[c]$ represents the rows selected from B corresponding to the real Fermion configuration.

The evaluation of both the "real" and "hidden" single particle orbitals on the extra sites is done implicitly. The NN outputs the value of the original N_s and extra r SPO's with respect to the (never explicitly specified) locations of the r hidden fermions on the hidden sites. This gives the $r \times N_s$ matrix $C_\theta(c)$ for the original SPO's and the $r \times r$ matrix $F_\theta(c)$ for the new SPO's.

This will then give an amplitude, for r hidden fermions of

$$\Psi_{r\text{-HF}}(c) = \det \begin{bmatrix} A[c] & B[c] \\ C_\theta(c) & F_\theta(c) \end{bmatrix}. \quad (6)$$

In the case of HFDS [Eq. (6)], the NN parameters θ and A and B are all optimized when applied to the ground state approximation for fermionic Hamiltonian.

IV. NUMERICAL METHODS

To support our statements about the relation among the NN wave functions, we will provide some numerical evidences. In this section, we will present the numerical methods for the ground state approximation of the Hubbard model, and the loss functions we define for mutual learning of wave

functions, and the quantities for measuring the modulation on the variational wave-function amplitudes.

A. Model and wave-function architecture

The variational wave functions are optimized to approximate the ground state of the Hubbard model

$$\hat{H} = -t \sum_{\langle i,j \rangle, \sigma} (\hat{c}_{i\sigma}^\dagger \hat{c}_{j\sigma} + \text{H.c.}) + U \sum_i \hat{n}_{i\uparrow} \hat{n}_{i\downarrow}, \quad (7)$$

where $\hat{c}_{i\sigma}$ ($\hat{c}_{i\sigma}^\dagger$) is annihilation (creation) operator for spin- σ ($\sigma = \uparrow, \downarrow$) electron on site i , and $\hat{n}_{i\sigma} = \hat{c}_{i\sigma}^\dagger \hat{c}_{i\sigma}$ is the number operator. We will consider this model on an $L = 4 \times 4$ square lattice with periodic boundary conditions on both spatial directions, interaction strength $U/t = 8$, and $\frac{1}{8}$ hole doping, i.e., $N_\uparrow = N_\downarrow = 7$ [26].

The neural network architecture is designed to have two hidden layers, each with 2048 neurons and ReLU activation function, parametrized by θ , the set of weights and bias inside the NN. The output is an oversized array, which is reshaped into an $r \times N_s$ matrix $C_\theta(c)$ and an $r \times r$ matrix $F_\theta(c)$, which, together with the static matrices (with entries independent of c) A and B , are used to construct the SPO and corresponding determinant matrices for various types of wave functions considered in this work [see Fig. 1(e)].

B. Ground state approximation

To optimize the wave functions towards the ground state, in this work, we use supervised wave function optimization with Adam [27,28] (SWO-Adam, see Appendix H). Within SWO, at each so-called imaginary-time step t , we set a target state $|\Phi_t\rangle = e^{-\tau \hat{H}} |\Psi_{t-1}\rangle \approx (1 - \tau \hat{H}) |\Psi_{t-1}\rangle$ for the present variational wave function $|\Psi_t\rangle$. And a loss function to minimize in this procedure is defined as the logarithm of the wave function fidelity:

$$\mathcal{L} = -\ln \frac{\langle \Psi_t | \Phi_t \rangle \langle \Phi_t | \Psi_t \rangle}{\langle \Psi_t | \Psi_t \rangle \langle \Phi_t | \Phi_t \rangle}. \quad (8)$$

The gradient of Eq. (8) is estimated with Monte Carlo sampling from probability distribution $p(c) = |\Psi_{t-1}(c)|^2$,

$$\frac{\partial \mathcal{L}}{\partial \theta} \approx 2 \sum_{\{c|p(c)\}} \Re \left[\alpha(c) \text{Tr} \left(\tilde{A}_\theta(c)^{-1} \frac{\partial \tilde{A}_\theta(c)}{\partial \theta} \right) \right], \quad (9)$$

where the coefficient $\alpha(c)$ is given by

$$\alpha(c) = \frac{|\psi(c)|^2}{\sum_{\{c\}} |\psi(c)|^2} - \frac{\psi(c)^* \phi(c)}{\sum_{\{c\}} \psi(c)^* \phi(c)}, \quad (10)$$

$$\psi(c) = \frac{\Psi_t(c)}{\Psi_{t-1}(c)}, \quad \phi(c) = \frac{\Phi_t(c)}{\Psi_{t-1}(c)}. \quad (11)$$

By iteratively feeding the estimated gradient Eq. (9) into Adam optimizer to minimize the loss function Eq. (8) at each step t , the variational state $|\Psi_t\rangle$ is undergoing imaginary time evolution stochastically, with energy being decreased until convergence.

The results, in terms of relative energy error to the exact ground energy for the chosen model $E_0 = -11.868$ [26], are shown in Fig. 2.

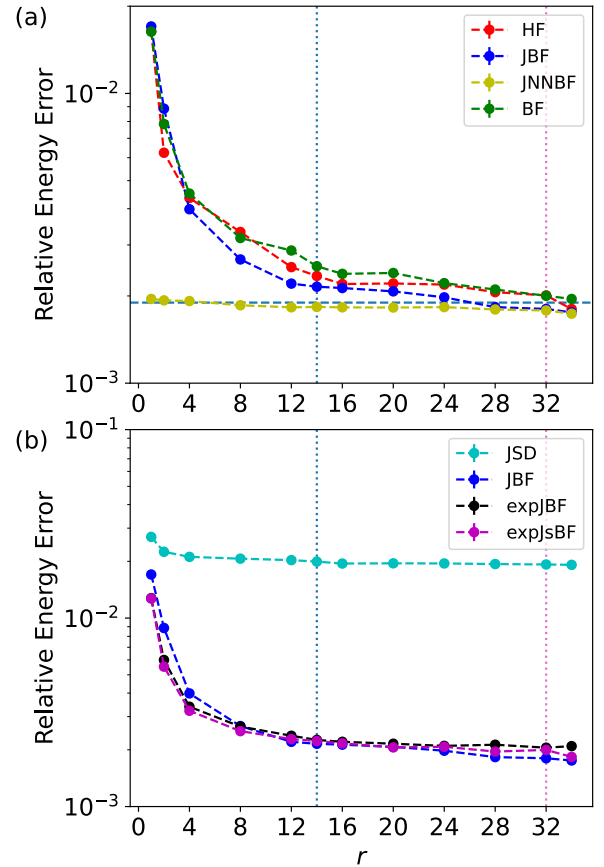


FIG. 2. Variational energy results from ground state approximation on the Hubbard model ($L = 4 \times 4$, PBC, $U/t = 8$, $N_\uparrow = N_\downarrow = 7$). The two vertical dotted line show the locations of $r = N_s$ (blue) and $r = L_s$ (pink), and the horizontal dashed line in (a) is the lowest energy from r -BF at $r = L_s$ from Fig. 6.

C. Mutual learning of wave functions

The mutual learning of wave functions is the optimization of one wave function $|\Psi_1\rangle$, e.g., r -JBF (see Sec. V A), towards the other one $|\Psi_2\rangle$, e.g., r -HF. Two metrics are defined for this purpose: one is the wave-function fidelity

$$\mathcal{F}_{\text{wfi}} = \frac{\langle \Psi_1 | \Psi_2 \rangle \langle \Psi_2 | \Psi_1 \rangle}{\langle \Psi_1 | \Psi_1 \rangle \langle \Psi_2 | \Psi_2 \rangle} \quad (12)$$

and the other is the square distance between the SPO's

$$\mathcal{L}_{\text{spo}} = \frac{1}{S} \sum_{\{c\}} |\tilde{A}_1(c) - \tilde{A}_2(c)|^2, \quad (13)$$

where $\tilde{A}_{1/2}(c)$ are the $L_s \times N_s$ -size SPO matrices generated from $|\Psi_{1/2}\rangle$ for configuration c , and S is the number of samples in $\{c\}$. This latter metric is only applicable when both ansatz generate SPO's. We also defined the SPO fidelity as

$$\mathcal{F}_{\text{spo}} = \sum_{\{c\}} \frac{\det(\mathcal{S}_1(c)^\dagger \mathcal{S}_2(c)) \det(\mathcal{S}_2(c)^\dagger \mathcal{S}_1(c))}{\det(\mathcal{S}_1(c)^\dagger \mathcal{S}_1(c)) \det(\mathcal{S}_2(c)^\dagger \mathcal{S}_2(c))} \quad (14)$$

to measure the average overlap between the SPO's for each configuration c .

These metrics are used as the distances between two wave functions, and the target states are those from the ground state approximation for Hubbard model.

D. Amplitude modulation

To give an estimate for how rapidly various wave functions oscillate both with respect to the magnitude and sign, we compute the following for different parts of the wave function $f(c)$.

Given a Markov chain for Monte Carlo sampling from wave function $|\Psi\rangle$, we collect all the configuration samples from this chain removing successive identical configurations, such that the remaining ones in the set $\{c_j\}_{j=1}^J$ have the neighboring samples differing by the position of exactly one fermion, i.e., $|c_j - c_{j+1}| = 2$ for all c_j . We then consider $\ln f_{\text{ratio}}(j) \equiv \ln f(c_{j+1}) - \ln f(c_j)$ with a particular focus on its average real and imaginary pieces, i.e.,

$$\mathcal{F}_{\text{magn}} = \frac{1}{J} \sum_{j=1}^J |\Re \ln f_{\text{ratio}}(j)|^2, \quad \mathcal{F}_{\text{phase}} = \frac{1}{J} \sum_{j=1}^J |\Im \ln f_{\text{ratio}}|^2, \quad (15)$$

which are taken over the sample set $\{c_j\}_{j=1}^J$. These quantities are independent of the norm of the wave function and allow separate estimates of the magnitude and phase oscillations.

V. COMPARISON BETWEEN WAVE FUNCTIONS

We start by rewriting HFDS with r hidden fermions in a somewhat different form. Without loss of generality, we assume that $F_\theta(c)$ is invertible for all c (see Appendix A). Then we can rewrite Eq. (6) as

$$\Psi_{r\text{-HF}}(c) = \det(F_\theta(c)) \det((A - BF_\theta(c)^{-1}C_\theta(c))[c]). \quad (16)$$

In the parlance of NNBF, the r -HF wave function is neural network backflow with an additive correction of the form

$$\tilde{A}_\theta(c) = A - B(F_\theta(c))^{-1}C_\theta(c) \quad (17)$$

multiplied by a determinant factor $F_\theta(c)$ (alternatively an additive correction composed with a right-rotation with F).

This formula can be derived either directly by working with determinants of block matrices; or alternatively by using the matrix determinant lemma and computing the change in the determinant between the (padded with identity) matrix $A[c]$ and the matrix $(A + BC_\theta(c))[c]$.

r -HF then is a form of NNBF but does differ from the most standard Jastrow-NNBF in two ways. First, there is a somewhat nonstandard ‘‘Jastrow factor’’ consisting of the determinant of a $r \times r$ matrix, which can be negative. We will want to address two questions about this determinant factor: (a) is it preferable to use a determinant factor compared to a more basic (possibly also NN-inspired) Jastrow; (b) does the addition of a determinant factor on top of NNBF span a larger class of wave functions than no multiplicative factor at all. Second, we will be interested in understanding the difference in the additive corrections between r -HF and standard NNBF. Allowing the NNBF to use as its multiplicative factor an $r \times r$ determinant and fixing the NN’s to the same depth and width, we want to ascertain whether using the NN output for matrices C and F which are multiplied as in Eq. (17) to generate the additive SPO is better than just outputting M directly.

A. Comparisons of additive corrections

We will address these in reverse order starting first with comparing the additive correction between r -HF and JNNBF.

To make progress on this, it will also be useful to define a closely related wave function

$$\tilde{A}_\theta(c) = A + BC_\theta(c) \quad (18)$$

as well as its determinant-factor multiplied form r -JBF (equivalently r -TBF). r -JBF differs from r -HF in that it directly uses the outputted matrix $C_\theta(c)$ whereas r -HF uses $(F_\theta(c)^{-1})C_\theta(c)$.

If we consider Eq. (17) or (18), we see that the additive correction to the SPO is the product of an $L_s \times r$ matrix (B) and an $r \times N_s$ matrix, $(F_\theta^{-1}C_\theta$ or C_θ), resulting in a low-rank additive correction of, at most, rank r . As generically the rank of the additive correction to JNNBF will be full rank (i.e., N_s), almost none of the JNNBF SPO’s will be representable by either Eq. (17) or (18) when $r < N_s$. On the other hand, we can show every additive SPO generated by r -JBF can also be generated by JNNBF. This is done by converting the r -JBF neural network (even for $r > L_s$) into a JNNBF neural network of the same size by first increasing the r -JBF neural network by one linear layer (encoding B in the weights) so that it outputs $M_\theta(c)$ and then compressing the two final linear layers into a single linear layer. This results in a JNNBF wave function with the same initial layers. Notice, it is also straightforward to see that L_s -JBF and JNNBF span the same space (let B be the identity) and that $(r+1)$ -JBF contains r -JBF (but the converse is generically not true when $r < N_s$). In Fig. 2, we also see evidence of this numerically as the energy of r -HF (or r -JBF) decrease monotonically with r out to, at least, $r = L_s$ with the most significant decreases at $r < N_s$. See Appendix B for more explicit arguments for these descriptions.

Analytically proving whether r -HF is strictly included in JNNBF is made difficult by the F^{-1} term. That said, JNNBF is almost certainly larger than r -HF (for $r \leq L_s$) as we expect on general grounds that r -JBF and r -HF span a very similar (if not identical) space. In fact, in the $w \rightarrow \infty$ limit they do span the same set of configuration-dependent SPO’s under the assumption of universal approximation theorem for arbitrary-width NN [29]. At finite w , we have numerically considered low-energy (with respect to the 4×4 Hubbard model) r -JBF and r -HF states. In Fig. 3, we show the fidelity after mutual learning between r -HF and r -JBF ($r \leq N_s$) by maximizing the wave-function fidelity Eq. (12). We found that r -HF and r -JBF can be optimized towards each other (in both directions) so that their infidelity is lower than 10^{-2} . This is consistent also with finding their optimized energies being very close with r -JBF being slightly lower other than at $r = 2$. Unsurprisingly, we find that JNNBF finds lower-energy variational states than both r -JBF and r -HF at all r , as shown in Fig. 2.

This leaves us with the following situation:

$$r\text{-HF} \approx r\text{-JBF}, \quad (19)$$

$$r\text{-JBF} \underset{\text{spo}}{\subset} (r+1)\text{-JBF} \underset{\text{spo}}{\subset} L_s\text{-JBF} = \text{JNNBF}, \quad (20)$$

$$r\text{-HF} \underset{\text{spo}}{\subset} \text{JNNBF} : w \rightarrow \infty, \quad (21)$$

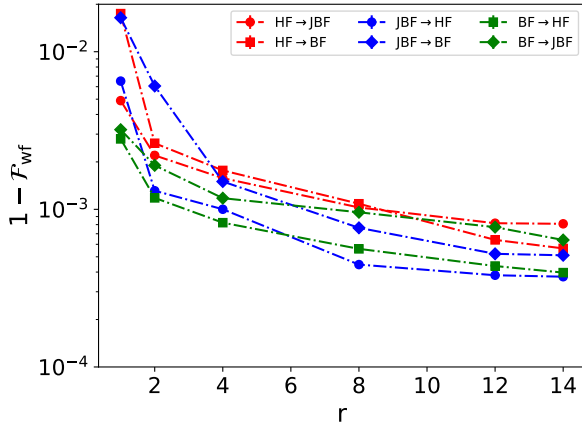


FIG. 3. Wave-function fidelity from mutual learning of different wave functions by *maximizing the wave-function fidelity*: optimization of r -HF towards r -JBF (red circle), and $2r$ -BF (red square); r -JBF towards r -HF (blue circle), and $2r$ -BF (blue diamond); $2r$ -BF towards r -HF (green square), and r -JBF (green diamond). The target states are from the low energy space of Hubbard model ($L = 4 \times 4$, PBC, $U/t = 8$, $N_\uparrow = N_\downarrow = 7$).

where we use the notation \subset_{spo} to indicate one class of wave functions contains the same set of SPO's as the other class (a strictly stronger inclusion than the wave functions being the same). Notice that Eq. (20) is also true if we remove the determinant factor from each term.

B. Benefit of multiplicative factor on neural network backflow

In the previous section, we argued that when dressed with a determinant factor, the additive corrections being used by JNNBF were generically a superset of the additive correction of other wave functions such as r -HF and r -JBF whose additive corrections had restricted rank. In this section, we will consider whether multiplying by a determinant out front—i.e., the determinant factor—is important at all—i.e., is NNBF identical, in some limits, to JNNBF.

To make progress on this, we will actually start by focusing on comparing r -JBF (or r -HF) with r -BF. To make this comparison, we will take the r -JBF (r -HF) wave functions and rewrite them as a BF wave function with a different additive correction—i.e., we will have the multiplicative factor absorbed into the additive correction.

When $r \leq N_s$, we can rewrite (see Appendix C for details), these wave function as

$$\Psi_{r\text{-HF/JBF}} = \det((A - B'C'_\theta(c))[c]), \quad (22)$$

where

$$B' = [-A_1 \quad B], \quad (23)$$

and

$$C'_\theta(c) = \begin{cases} \begin{bmatrix} F - I & \mathbf{0} \\ F^{-1}C_1F & F^{-1}C_2 \end{bmatrix}, & \text{for } r\text{-HF} \\ \begin{bmatrix} F - I & \mathbf{0} \\ C_1F & C_2 \end{bmatrix}, & \text{for } r\text{-JBF} \end{cases} \quad (24)$$

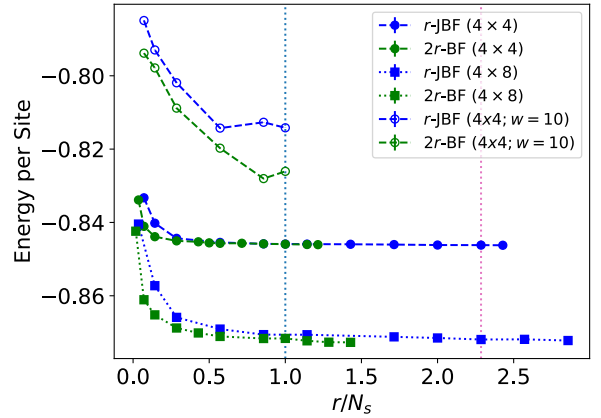


FIG. 4. Replot of the energy data from Fig. 2 for comparison between r -JBF (blue) and $2r$ -BF (green). In addition to the results for previous model ($L = 4 \times 4$, PBC, $U/t = 8$, $N_\uparrow = N_\downarrow = 7$) (circle markers), the results from a larger lattice ($L = 4 \times 8$, PBC, $U/t = 8$, $N_\uparrow = N_\downarrow = 14$) (square markers), and results for small-size NN (width=10) (empty circle markers) are displayed as well.

are of size $L_s \times 2r$ and $2r \times N_s$, respectively. Notice that this means each r -JBF (r -HF) wave function can be written as a slightly more complicated additive correction of rank $2r$. As both $2r$ -BF and r -JBF can be written as a neural network backflow with rank- $2r$ additive corrections, this suggests that they might span a similar space of wave functions. Generically, we might expect that $2r$ -BF actually spans a larger space because the $2r \times L_s$ SPO's are completely general for $2r$ -BF while for r -JBF (r -HF) the upper right $r \times (N_s - r)$ corner of the additive correction is forced to be zero. In the $w \rightarrow \infty$ limit, it is strictly true that $2r$ -BF contains the (determinant absorbed) SPO's of both r -JBF and r -HF (but not the converse).

Unsurprisingly, we find that energetics of $2r$ -BF and r -JBF are similar on both 4×8 and 4×4 Hubbard models (see Fig. 4) with $2r$ -BF being slightly lower in energy.

Moreover, through numerical fidelity matching of the low energy states (see Fig. 3), we find that $2r$ -BF can represent both r -JBF and r -HF with infidelities less than 10^{-2} for all r . In the case of r -JBF, essentially this tells us that $2r$ -BF can output the matrix C_1F from r -JBF where previously it was only outputting F and C separately. The converse (r -JBF and r -HF matching $2r$ -BF) also achieves infidelities less than 10^{-2} except at $r = 1$ where the fidelity matching is worse in that direction.

We even attempt to go beyond just fidelity matching and try to directly optimize the SPO distance between $2r$ -BF and r -JBF by optimizing the $2r$ -BF wave function (the opposite is not possible because of the zeros). In Fig. 5, we also show the results from mutual learning by minimizing the SPO distance Eq. (13). Notice that optimizing SPO distance is more strict than optimizing the wave-function fidelity, as there is a significant amount of degree of freedom for the SPO to give the same amplitude during the operations of the row selection and determinant evaluation. Hence, in Figs. 5(b) and 5(c), we further show the SPO fidelity and the wave-function fidelity that are consistent with the SPO distance in Fig. 5(a). In some limit, optimizing the SPO distance and the fidelity are closely

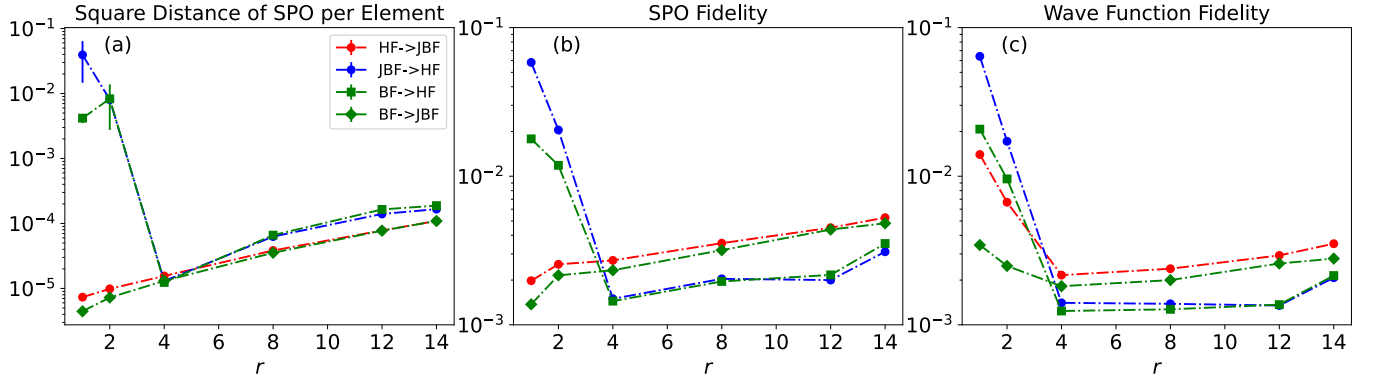


FIG. 5. (a) Square distance of SPO's per configuration and matrix element, (b) SPO fidelity ($1 - \mathcal{F}_{\text{spo}}$ in the plot) and (c) wave-function fidelity ($1 - \mathcal{F}_{\text{wf}}$ in the plot) from mutual learning of different wave functions by *minimizing the SPO distance*: optimization of r -HF towards r -JBF (red circle); r -JBF towards r -HF (blue circle); $2r$ -BF towards r -HF (green square), and r -JBF (green diamond). The target states are from the ground state approximation for Hubbard model ($L = 4 \times 4$, PBC, $U/t = 8$, $N_{\uparrow} = N_{\downarrow} = 7$).

related (i.e., their respective infidelities become zero at the same time), but they do probe the closeness of wave functions in different ways. For example, it is possible to have low SPO-distance but not super-high fidelity because even if the SPO distance is small, its contribution to the fidelity can be amplified by large values generated from the inverse of the row-selected SPO (see Appendix D for a detailed discussion of this). We find that the $2r$ -BF can match the SPO's quite successfully at all r 's for r -JBF and at $r \geq 4$ for r -HF.

While the above numerical experiments are performed by fixing the NN for all the wave functions, we can also ask whether there are “minor” architectural changes to the FFNN for $2r$ -BF which would allow for an exact reconstruction of other wave functions. We demonstrate an example of this reconstruction in Appendix E where we write a r -JBF wave function as a r -BF wave function with two additional layers (with different activation functions) and widths expanded only by an extra factor of r .

It is worth mentioning that it is also possible to reconstruct r -HF from $2r$ -BF by adding a few more layers to the FFNN, where we can use adjugate formula to obtain the inverse of F , i.e., $F^{-1} = \text{adj}(F)/\det F$. However, this involves the evaluation of determinants within the NN, and it will require widths on the order of $r!$ number of neurons if implemented directly via the definition of determinant making this direct approach impractical (we know that the evaluation of determinant can be done in $O(n^3)$ steps, but this will require many more layers within the NN). Nevertheless, the fidelity and the energetics from Fig. 2 indicate that even without any change in architecture, r -JBF or NNBF are already as expressive as r -HF and effective in approximating the ground state without the necessity of F^{-1} to be included in the additive correction.

We now directly consider the relation between NNBF and r -JBF/ r -HF. On the one hand, when $2r < L_s$, almost none of the NNBF SPO's will be representable by either of these two wave functions. On the other hand, we already know that all the $2r$ -BF wave functions are representable by NNBF and given the evidence that $2r$ -BF is a superset of r -JBF, this suggests that all ($r \leq N_s$) r -JBF have (determinant absorbed) SPO's that are representable by NNBF; this is rigorously true for $w \rightarrow \infty$. This argument also suggests that JNNBF and NNBF should span the same space for all $r \leq N_s$. Upon

optimization (see Fig. 2), we find that their energetics are extremely close with very minimal r dependence in JNNBF out to $r = N_s$ (surprisingly actually even for $r > N_s$ there is still very little r dependence). We again attempt to match the fidelity of r -JNNBF with NNBF; This optimization is done through L_s -BF which is equivalent to NNBF. In Fig. 6, we show the fidelity after learning r -JNNBF ($r \leq N_s$) with L_s -BF, and the comparison between energy after the optimization.

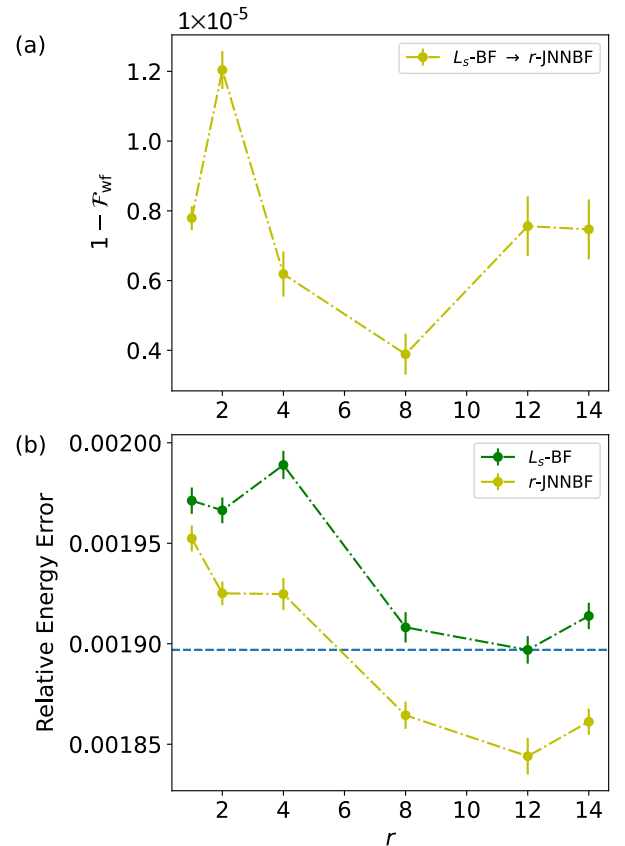


FIG. 6. Results of mutual learning of L_s -BF towards r -JNNBF by *maximizing the wave-function fidelity*. (a) Wave-function fidelity (b) variational energy after the optimization. The horizontal lines in (b) is the lowest energy of L_s -BF after the optimization.

And we find an infidelity of essentially 10^{-5} for all r , suggesting that NNBF really does contain JNNBF.

The conclusion of this section is that including a determinant factor which multiplies an NNBF wave function seems to not increase the space of wave functions spanned by NNBF, and increases the space of wave functions spanned by r -BF up to $2r$ -BF when $r \leq N_s$. When the number of hidden fermions is very large - more than the number of actual fermions, it is less clear whether the determinant factor should have additional benefit although in practice we see almost no decrease in energy for JNNBF at $r > N_s$ (see Fig. 2).

C. Exponential Jastrow versus determinant factor

What we have seen in the previous section, is that the determinant factor matters very little for JNNBF, but does increase the “rank” or accessible matrix size of BF from r to $2r$. Also, when $r > N_s$, formally the determinant factor can be important. If one is going to use a multiplicative factor to enhance the wave function, it is worth considering whether the determinant factor is better than a more standard alternative of a Jastrow factor. In particular, a reasonable comparison is to use an exponential Jastrow where the term being exponentiated also comes from NN; we call this class of wave functions expJBF. Like a standard Jastrow, we have that there is always a positive correction to the determinant backflow. We can additionally test whether this is important by one more multiplicative factor, expJsBF which multiplies the Jastrow factor by an additional NN-generated scalar that is allowed to be positive and negative.

In this section, we primarily compare the energetic of these wave functions, to numerically check which have lower energy on our prototypical 4×4 Hubbard model [see Fig. 2(b)]. We find that the energy of expJBF and expJsBF are almost identical although the signed version gives slightly better results at large r . In comparison against JBF, we find that at small rank, we get lower energy from the exponential Jastrows than the determinant factor; this relationship seems to switch at large enough r .

A plausible mechanism for this difference is that the role of the Jastrow or the multiplicative factor is to be able to introduce exponentially large separation between different configurations. An exponential Jastrow naturally has this ability; the determinant factor essentially emulates this exponential separation through the product of the respective matrices eigenvalues. At low r , it is restricted in the number of such terms in the product and has trouble matching the exponential Jastrow. This all said, the difference between the various multiplicative factors are tiny especially when r is large, and it is not clear that one should take those tiny differences at this level seriously. The high-level result is that some multiplicative factor marginally improves a restricted form of the NNBF wave function, but different reasonable multiplicative factors lead to essentially the same marginal improvement; there seems to be nothing fundamentally special or powerful about the determinant factor.

D. Determinant factor versus effective determinant factor

Throughout this paper, we have been making a somewhat artificial distinction between the determinant factor and Slater

determinant pieces of the wave functions. There is a sense in which one could make a different decomposition of the wave functions we have been considering, by writing the wave function as a fixed Slater determinant by an additional “effective determinant factor.” In particular, we can write all the wave functions considered so far as

$$\Psi(c) = \det(A[c]) \det(\bar{F}(c)), \quad (25)$$

where \bar{F} absorbs the rest of the wave function except the static Slater determinant part $A[c]$. Notice that (1) \bar{F} will depend on the NN in some way and (2) unlike all the other multiplicative factors we have considered, the matrix \bar{F} for which we are taking the determinant of may involve row selection. In particular, we can write NNBF as

$$\Psi_{\text{NNBF}}(c) = \det(A[c]) \det(I - CA[c]^{-1}B[c]) \quad (26)$$

and r -HF as

$$\Psi_{r\text{-HF}}(c) = \det(A[c]) \det(F - CA[c]^{-1}B[c]). \quad (27)$$

Naively, it is not obvious why (or if) these wave functions should be superior to a Slater determinant dressed with a determinant factor, i.e., $\bar{F} = F_\theta(c)$ which we call JSD. Numerically, we have examined this wave function and find JSD is much worse in energy than any of the NNBF-type wave functions. It is naively surprising that the NN in JSD is unable to learn $F_\theta(c) = \det(I - CA[c]^{-1}B[c])$ at our given width; notice in the $w \rightarrow \infty$ limit JSD could learn that $F_\theta(c)$.

To understand the mechanism for this distinction, let us first look at the determinant factor $\det F(c)$. The NN, as well as the determinant function, is a smooth function where small change of the input value will only cause a small change to the output; the determinant factor will not change drastically as the input c moves between the nearest configurations. However, the amplitude of the exact ground state is not necessarily a smooth function of the configurations, thus is likely to go beyond the modulation ability of a determinant factor at moderate w .

On the other hand, the effective determinant factor from NNBF $\det(I - CA[c]^{-1}B[c])$, has the presence of $A[c]^{-1}B[c]$, which is essentially not a smooth function of input configurations due to row selection and will impose a drastic change on top of the smooth change from the NN, making it possible to capture the pattern from the true ground state, while still maintaining the generalization ability of the NN even at small w .

We can explicitly test and quantify whether the effective determinant factor with row selection generates significantly stronger modulation than directly using the NN output as a determinant factor. To accomplish this, we want to look at the difference in amplitudes for configurations that are “nearest neighbors” (see Sec. IV D for details).

Here the following quantities will be considered: (1) $f(c) = \det F_\theta(c)$, where $\det F_\theta(c)$ is the determinant factor from the definition of wave-function ansatz, e.g., the one from Eq. (16); (2) $f(c) = \det \bar{F}(c)$, the effective multiplicative factor obtained by factoring out the static Slater determinant part of the wave function, e.g., $\det \bar{F}(c) = \det(F_\theta(c) - C_\theta(c)A[c]^{-1}B[c])$ for r -HF; (3) $f(c) = \mathcal{S}_{k_j}(c)$, the matrix element from the SPO.

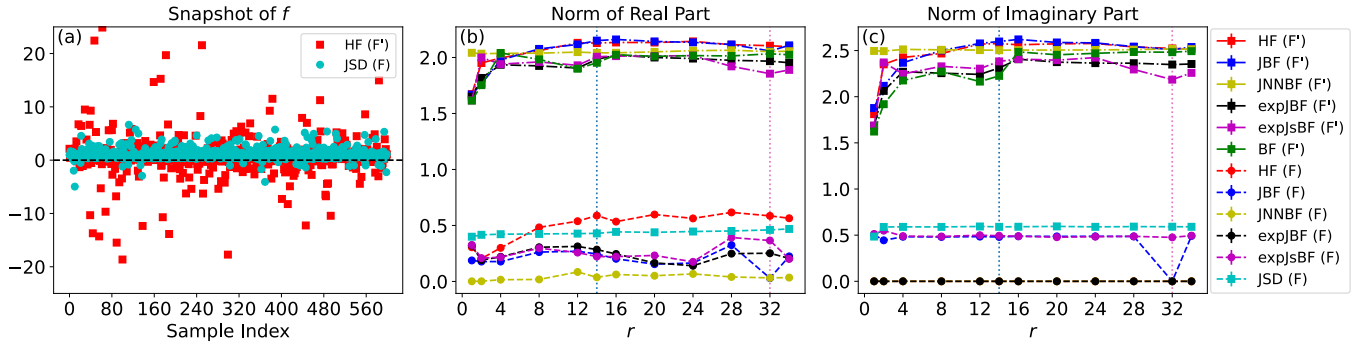


FIG. 7. (a) Samples of f_{ratio} for the effective determinant factor from r -HF and r -JSD ($r = N_s$) along a Markov chain. (b) The average square norm of the real part of f' . (c) The average square norm of the imaginary part of f' . In (b) and (c), the results for the defined multiplicative factors (F) and the effective multiplicative factors (F') are plotted for all the wave functions considered in this paper. The wave functions are from the ground state approximation for Hubbard model ($L = 4 \times 4$, PBC, $U/t = 8$, $N_\uparrow = N_\downarrow = 7$).

First, we look at a prototypical example for \bar{F} of both r -HF and r -JSD at $r = N_s$ where we measure the ratio of nearest-neighbor amplitudes of the effective determinant factor, $f_{\text{ratio}}(c_j) = \det \bar{F}(c_{j+1}) / \det \bar{F}(c_j)$ for random configurations sampled from $|\Psi(c)|^2$.

In Fig. 7(a), we can readily see an obvious distinction between the f_{ratio} , from r -HF and r -JSD. More specifically, the ratios from r -JSD show that the determinant factor, $\det \bar{F} = \det F_\theta$, modulates the wave-function amplitude in a weak way for both the sign structure and the magnitude. In contrast, the effective determinant factor from r -HF modulates both the magnitude and the sign structure in a strong way.

We can further quantify this modulation by computing the real and imaginary components of the log-ratio $\ln f_{\text{ratio}}(c_j)$ which act as an effective derivative, and the norms of them measure the modulation on the magnitude and sign structure, respectively. Using the logarithm importantly lets us see rapid relative changes in the amplitude even when $|f_{\text{ratio}}| \ll 1$. In Figs. 7(b) and 7(c), we again observe a sharp distinction between the effective determinant factors and determinant factors from all types of the wave functions considered in this work. Note that the larger the norm is, the stronger the modulation is, so that the modulation from effective determinant factor is nearly $\exp(0.75) \approx 2.1$ times stronger, based on the differences of the average norms in Figs. 7(b) and 7(c), than the determinant factor on both magnitude and sign structure. Interestingly, we can also observe from Figs. 7(b) and 7(c) that at lower rank, the effective determinant factor modulations from r -HF, r -JBF and r -BF are weaker than their counterparts at higher rank, which is consistent with their relatively higher energies in Fig. 2.

Alternatively, we compute the log-ratio of the matrix elements from SPO, and show the mean and variance over these $L_s \times N_s$ elements in Fig. 8. From them, we can extract some information about the modulation on the SPO from various types of wave-function ansatz. Although the construction of SPO does not involve the row selection as the wave-function amplitude, we still observe weaker modulations from JSD. The mean for JSD is always lower than those from NNBF-type wave functions, in both magnitude (real part of log-ratio) and sign (imaginary part of log-ratio) modulation, indicating that on each SPO element, JSD changes the values in a

relatively weaker way. Moreover, the smaller variance from JSD shows that JSD impose a more uniform modulation over the entire SPO, suggesting that JSD is less capable of identifying which part of SPO is worth stronger modulation than the rest.

Considering the geometry for the SPO, it is easy to see that, in the cases of $r \leq N_s$, the SPO of JSD is always within the linear space spanned by the column vectors from the static matrix A , while NNBF is dynamically changing the space for the SPO given different configuration inputs. In Appendix G, we demonstrate that it is impossible for JSD to reproduce the SPO from NNBF.

Overall we see that the effective determinant factor with row selection imposes stronger modulation on the wave function amplitudes, achieving better performances on ground state approximation. The investigation on SPO also suggests that NNBF produces a broader class of SPO that goes beyond the representability of Slater-Jastrow type wave functions, indicating that NNBF is likely to be a better ansatz candidate for Fermionic systems.

VI. SUMMARY AND OVERVIEW

In this work, we have considered a series of different NNBF wave functions and their relation with HFDS. We show that HFDS can be written in the NNBF form with an $r \times r$ determinant factor, $\det F_\theta(c)$, and an additive correction to the SPO of the form $BF_\theta(c)^{-1}C_\theta(c)$, where $F^{-1}C$ is an $r \times N_s$ matrix [see Fig. 1(a) bottom and Eq. (16)]. This is not of full rank for $r \leq N_s$ and is to be contrasted with the standard NNBF where the full-rank additive correction (an $L_s \times N_s$ matrix) is generated directly.

The determinant factor piece turns out to not be critical for NNBF wave functions: it can be replaced by an exponential Jastrow which achieves similar (and often better) energetics [see Fig. 2(b)] or alternatively when $r \leq N_s$ it can be absorbed into the additive correction to the SPO resulting in a new more complicated additive correction BC' , where C' is now “taller” by size r [see right-hand side of Figs. 1(a) and 1(b)] and Eqs. (22)–(24).

We are then left comparing various different additive corrections of the form BC' where C' is generated from

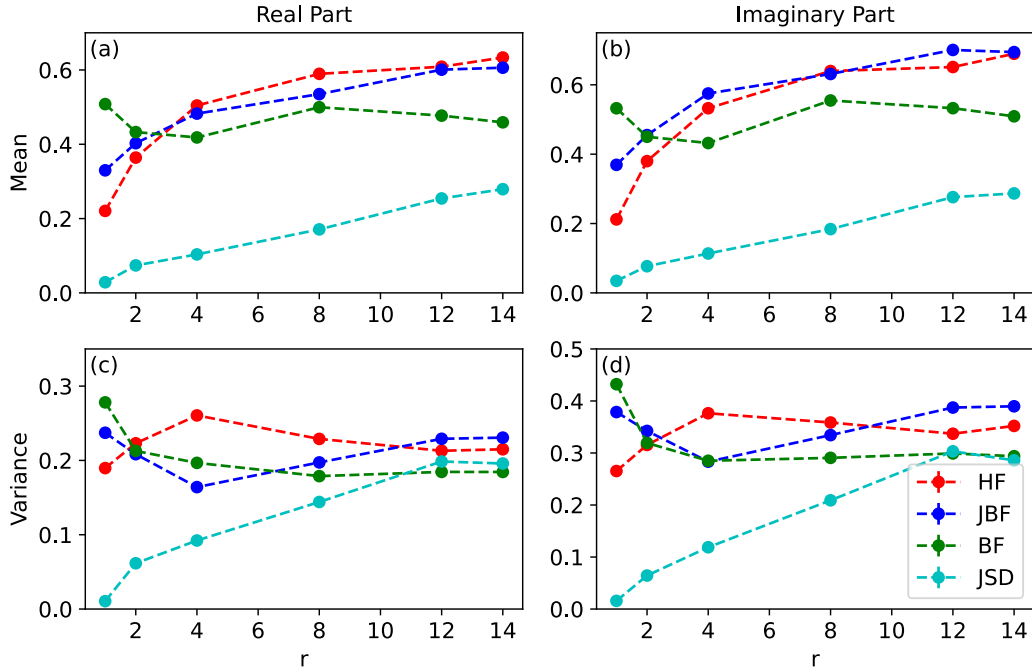


FIG. 8. The average norm of effective derivative on the SPO elements for r -HF, r -JBF, $2r$ -BF and r -JSD. (a) The mean of $\mathcal{F}_{\text{magn}}$ over all the SPO elements; (b) the mean of $\mathcal{F}_{\text{phase}}$; (c) the variance of $\mathcal{F}_{\text{magn}}$ over all the SPO elements; (d) the variance of $\mathcal{F}_{\text{phase}}$; and the wave functions are from the ground state approximation for Hubbard model ($L = 4 \times 4$, PBC, $U/t = 8$, $N_{\uparrow} = N_{\downarrow} = 7$).

various combinations of neural network output. Two general principles appear here. First and unsurprisingly, larger r is better until $r = L_s$. Larger r allows access to more flexibility in the matrices that can be generated including, out to $r = N_s$, access to higher rank matrices. Beyond $r \geq L_s$, we can reduce additive corrections to $r = L_s$. Secondly, the different ways we have considered to use the NN output to produce a C' of size $2r$ (i.e., r -JBF, r -HF, $2r$ -BF, etc.) all are roughly similar in energetics and expressibility and, where they differ, favor the states which are simpler, more direct, and more generic updates (i.e., $2r$ -BF). In particular, we see that r -HF and r -JBF are empirically very close energetically [with a slight edge to r -JBF, see Fig. 2(a)] and that we can optimize them towards each other with relatively low infidelity in each direction (see Fig. 3). $2r$ -BF is consistently lower in energy than both r -JBF and r -HF (see Fig. 4) and optimizing toward and from $2r$ -BF can be done with low infidelity (see Fig. 3). In the limit of large width, $2r$ -BF can represent any SPO in r -JBF and r -HF; and with two extra layers can represent any SPO in r -JBF at finite width. Amongst the various NNBF wave functions, these two principals suggest that using standard NNBF is better energetically and likely spans the space of all the other wave functions we have considered when $r \leq N_s$. For $r > N_s$, it likely makes sense to use NNBF with either an exponential Jastrow or determinant factor (although the improvement in energy even then is somewhat marginal).

NNBF and HFDS have both been successful in capturing highly accurate ground states. We compare a wave function JSD which is similar on the surface, a determinant factor times a Slater determinant, and find that it is significantly worse. To help understand what aspects are endowing NNBF with its particular advantage, we rewrite all our wave function as a (nonconfiguration independent) Slater determinant times an

“effective determinant factor” and then compare properties of these effective factors. In cases, such as NNBF and HFDS where the effective determinant factor involves row selection, we see that both the sign and amplitude are rapidly oscillating as a function of configuration, something which does not happen at moderate width for JSD. This suggests that some of the advantage of NNBF is coming from the ability to rapidly change the wave function.

Unlike the tensor network ansatz, which are well understood in terms of entanglement [30–35], the development of neural quantum states (NQS) largely relies on heuristic experiments, and lacks a comprehensive understanding for the underlying principles [23,36–41]. We hope our work paves the way for unifying various types of NQS and identifying the crucial components for an efficient wave function ansatz, which, in turn, may lead to the discovery of more accurate and scalable NQS.

ACKNOWLEDGMENTS

We acknowledge the helpful conversations with A. Liu, D. Luo, and K. Loehr. This work made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program (ICCP) in conjunction with the National Center for Supercomputing Applications (NCSA) and which is supported by funds from the University of Illinois at Urbana-Champaign. We acknowledge support from the NSF Quantum Leap Challenge Institute for Hybrid Quantum Architectures and Networks (NSF Award 2016136).

APPENDIX A: VALIDITY OF EQ. (16)

In Sec. V, we write down the transformed wave-function expression for HFDS, Eq. (16), by assuming that the backflow

matrix $F_\theta(c)$ is invertible. Here we are proving that even if originally $F_\theta(c)$ is not invertible for all the configuration input c 's, we can always construct an exact transformation (without changing the amplitude output or the neural network architecture) on HFDS, such that

(1) $F_\theta(c)$ is invertible for all c 's when the backflow matrix $[C_\theta(c) \ F_\theta(c)]$ for r -HF is at least rank r .

(2) Equation (16) is still valid for all c 's when there exists backflow matrix $[C_\theta(c) \ F_\theta(c)]$ for r -HF with rank less than r .

Given the original HFDS wave-function form

$$\Psi_{r\text{-HF}}(c) = \det \begin{bmatrix} A[c] & B[c] \\ C_\theta(c) & F_\theta(c) \end{bmatrix}. \quad (\text{A1})$$

We can divide the configuration basis sets into two disjoint sets: invertible set V whose elements c give invertible $F_\theta(c)$, and uninvertible set \bar{V} with uninvertible $F_\theta(c)$, or equivalently, $\det F_\theta(c) = 0$. The underlying reason for $\det F_\theta(c) = 0$ is that the $r \times r$ matrix $F_\theta(c)$ is made up of less than r , say, r_c , linearly independent vectors. Performing Gaussian elimination on the columns of $F_\theta(c)$, we will end up with

$$F_\theta(c) \rightarrow [\tilde{f}_{1,\theta}(c) \ \tilde{f}_{2,\theta}(c) \ \cdots \ \tilde{f}_{r_c,\theta}(c) \ 0 \ \cdots \ 0], \quad (\text{A2})$$

where $\{\tilde{f}_{j,\theta}(c)\}_{j=1}^{r_c}$ are r -dimensional column vectors. Because of the presence of zero vectors in Eq. (A2), we get the unwanted $\det F_\theta(c) = 0$ and uninvertible $F_\theta(c)$. Nevertheless, suppose that a $c \in \bar{V}$ has backflow matrix $[C_\theta(c) \ F_\theta(c)]$ is of rank at least r , but unfortunately has rank-deficient $F_\theta(c)$ as given by Eq. (A2), we can mix it with new linearly independent vectors from $C_\theta(c)$ to remove those zero vectors, such that $F_\theta(c)$ is full-rank and invertible. To do so, we introduce an $(N_s + r) \times (N_s + r)$ mixing matrix K , which is upper-triangular with 1's on the diagonal:

$$K = \begin{bmatrix} 1 & k_{12} & \cdots & k_{1,N_s+r} \\ 0 & 1 & \cdots & k_{2,N_s+r} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \quad (\text{A3})$$

We right multiply the r -HF matrix with this mixing matrix K , the result of which will be transparent if we rewrite the r -HF matrix in terms of column vectors:

$$\begin{aligned} & [A[c] \ B[c]C_\theta(c) \ F_\theta(c)]K \\ &= [h_1 \ \cdots \ h_{N_s} \ f_1 \ \cdots \ f_r]K \\ &= [\bar{h}_1 \ \cdots \ \bar{h}_{N_s} \ \bar{f}_1 \ \cdots \ \bar{f}_r], \end{aligned} \quad (\text{A4})$$

where

$$\bar{h}_m = \sum_{j=1}^{m-1} k_{j,m} h_j + h_m, \quad (\text{A5})$$

$$\bar{f}_m = \sum_{j=1}^{N_s} k_{j,N_s+m} h_m + \sum_{j=1}^{m-1} k_{N_s+j,N_s+m} f_j + f_m. \quad (\text{A6})$$

Effectively, the K matrix adds to each column vector a linear combination of previous ones. By choosing appropriate matrix entries for K , the right multiplication of K would increase the rank of $F_\theta(c)$ up to r (full-rank) via adding new column vectors from $C_\theta(c)$. As long as those $c \in \bar{V}$ satisfy the condition that the backflow matrix $[C_\theta(c) \ F_\theta(c)]$ is of rank at least r , there always exists some column vector from $C_\theta(c)$ to mix in to make $F_\theta(c)$ full-rank.

Moreover, as K is an upper-triangular square matrix with constant entries (not dependent on c), and $\det K = 1$, this mixing will not change the wave-function amplitude of HFDS. The practical realization on HFDS for this mixing is simply to add one more linear layer with weights constructed from K to the output layer of the NN that generates the backflow matrix. This new linear layer can just be absorbed into the last layer of the original NN, as a result, this mixing can be achieved by simply adjusting the weights and bias on the original output layer even without changing the NN architecture.

Since the effect of the mixing matrix K is to mix new vectors into $F_\theta(c)$, the choices for the nonzero entries of K are almost generic, as long as it circumvents the situation where the mixing accidentally decreases the rank for those originally full-ranked $F_\theta(c)$, $c \in V$. But this only happens for a zero-measure set in the parameter space defining K , which can be excluded from the determination of K .

At last, we consider the situation where the backflow matrix has rank less than r , which indicates that there are not sufficient column vectors from $C_\theta(c)$ to make $F_\theta(c)$ full-rank, such that $F_\theta(c)$ is always rank-deficient, i.e., $\det F_\theta(c) = 0$ no matter what mixing matrix K we have chosen. In this case, not only the columns of the $r \times (N_s + r)$ backflow matrix are linearly dependent, but the rows are also linearly dependent. As a consequence, the r -HF matrix for HFDS is rank-deficient, and will give zero amplitude from its determinant. Nevertheless, Eq. (16) is still valid, since it also gives zero amplitude as the original HFDS amplitude, due to $\det F_\theta(c) = 0$. Although it brings with an ill-defined F^{-1} in the expression, in practice, this none-valued output can be replaced with 0 during programming without affecting the performance.

In fact, this mixing procedure is also able to bring $F_\theta(c)$ away from rank-deficient regimes of θ , where $F_\theta(c)$ has extremely small but nonzero singular values and F^{-1} has extremely large entries, via, for example, optimizing the entries of matrix K with some proper loss function. By doing so, the learning of r -HF with r -JBF or r -BF will be practically easier.

APPENDIX B: EXPLICIT CONSTRUCTION FOR SUBSET SPO

In Sec. VA, we give statements on the relation between SPO's from different classes of wave-function ansatz. Here we provide explicit construction of SPO to support some of these statements.

(1) We start with r -HF $\subset (r+1)$ -HF for any r . Recall

that the SPO for r -HF is given by $A - BF^{-1}C$, which is an $L_s \times N_s$ matrix; it is straightforward to see that it can be

reproduced in the class of $(r + 1)$ -HF, by letting

$$A' = A, \quad B' = \left[\begin{array}{c|c} & \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix} \\ \hline B & \end{array} \right], \quad (\text{B1})$$

$$C' = \left[\begin{array}{c|c} C & \\ \hline 0 & 0 \dots 0 \end{array} \right],$$

$$F' = \left[\begin{array}{c|c} & \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix} \\ \hline F & \\ \hline 0 & 0 \dots 0 & | & 1 \end{array} \right], \quad (\text{B2})$$

such that $A' - B'F'^{-1}C' = A - BF^{-1}C$ for any c . Note that the constant elements on the backflow matrices C and F can be constructed by assigning the weights on the final linear layer to those positions as 0 and the bias as the corresponding constant values (0 or 1 here).

Similarly, we can also show that $r\text{-JBF} \subset_{\text{spo}} (r + 1)\text{-JBF}$ and $r\text{-BF} \subset_{\text{spo}} (r + 1)\text{-BF}$, using the same construction as $r\text{-HF}$ (Eqs. (B1) and (B2)), such that $A' - B'C' = A - BC$ for any c .

(2) Next, we show that for any r , including the cases of $r \geq L_s$, we have $r\text{-JBF} \subset_{\text{spo}} L_s\text{-JBF}$ and $r\text{-BF} \subset_{\text{spo}} L_s\text{-BF}$, both of which amount to showing $A' - B'C' = A - BC$, where B', C' are now matrices of size $L_s \times L_s$ and $L_s \times N_s$, respectively. Copy the neural network generating C for $r\text{-BF}$ or $r\text{-JBF}$ and add one more linear layer on top of it. The weights and bias of this additional layer are given by

$$W_{(mn)(kl)} = \delta_{nl} B_{mk}, \quad b_{(mn)} = 0, \quad (\text{B3})$$

where we use the matrix index (mn) , (kl) to index the weights and bias. The output of this new linear layer will directly give us the $L_s \times N_s$ matrix $C' = BC$, because

$$C'_{mn} = \sum_{k,l} W_{(mn)(kl)} C_{kl} + b_{(mn)} = \sum_k B_{mk} C_{kn} \equiv (BC)_{mn}. \quad (\text{B4})$$

Note that this additional linear layer (W', b') can be absorbed into previous linear layer (W, b) for C simply by $W \leftarrow W'W$, $b \leftarrow W'b + b'$. As we have gotten $C' = BC$, this construction is completed by the assignments of $A' = A$ and $B' = I_{L_s}$.

(3) Here we demonstrate that, given an NNBF that outputs matrix M [see Eq. (2)], we can generate an $L_s\text{-BF}$, that outputs matrix C , such that $BC = M$.

We start by rewriting a general $L_s\text{-BF}$ where we absorb B into the NN, which amounts to adding one more linear layer with weights as

$$W' = \left[\begin{array}{c|c|c|c} B & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{0} & B & \dots & \mathbf{0} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{0} & \mathbf{0} & \dots & B \end{array} \right], \quad (\text{B5})$$

where this is an $N_s \times N_s$ block matrix; We can combine it with the weight matrix from previous linear layer W which is

given as

$$W = \left[\begin{array}{c} W_1 \\ W_2 \\ \vdots \\ W_{N_s} \end{array} \right], \quad (\text{B6})$$

where W_h , $h = 1, 2, \dots, N_s$ (H is the width of last hidden layer) are arbitrary $r \times H$ matrices. Now the resulting weight $W'' = W'W$ is still a generic $(L_s N_s) \times H$ matrix with no constraints, because each block of W'' is given by

$$W''_h = BW_h \quad (\text{B7})$$

and is of full-rank L_s as B and W_h are of size $L_s \times r$ and $r \times H$, respectively, with $r \geq L_s$. We can assume B is of full rank L_s (if not, it can be brought to full rank by mixing with A , see Appendix A), such that there are complete sets of linearly independent vectors within it. Finally, the arbitrary choice of W_h enables W''_h to be any $L_s \times H$ matrix.

The reverse direction of the relation between NNBF and $r\text{-BF}$ ($r \geq L_s$) is also true by simply choosing a common basis set as B for each W_h .

APPENDIX C: ABSORPTION OF DETERMINANT FACTOR INTO SPO

In this section, we show the derivation for the absorption of determinant factor into SPO for $r\text{-HF}$ and $r\text{-JBF}$.

For $r \leq N_s$, we embed the $r \times r$ matrix $F_\theta(c)$ in an $N_s \times N_s$ block-diagonal matrix with identity on the extra diagonal block

$$[F_\theta] \rightarrow \left[\begin{array}{c|c} F_\theta & \mathbf{0} \\ \hline \mathbf{0} & I_{N_s-r} \end{array} \right] \equiv \tilde{F}_\theta. \quad (\text{C1})$$

This leaves the wave-function amplitude unchanged, since $\det F = \det \tilde{F}$, but allows us to multiply the two square matrices from the determinant factor and from the SPO's together before the determinants are computed

$$\begin{aligned} \Psi_{r\text{-HF}}(c) &= \det F_\theta(c) \det((A - BF_\theta^{-1}(c)C_\theta(c))[c]) \\ &= \det \tilde{F}_\theta(c) \det((A - BF_\theta^{-1}(c)C_\theta(c))[c]) \\ &= \det[(A - BF^{-1}C)\tilde{F}[c]] \\ &= \det[(A + A(\tilde{F} - I) - BF^{-1}C\tilde{F})[c]]. \end{aligned} \quad (\text{C2})$$

From Eq. (C2), we see a new SPO matrix of size $L_s \times N_s$. Next we transform it into BF.

After writing C in terms of block matrices as well, i.e.,

$$C = [C_1 \mid C_2] \quad (\text{C3})$$

where C_1 and C_2 are of size $r \times r$ and $r \times (N_s - r)$, respectively, we have

$$\begin{aligned} BF^{-1}C\tilde{F} &= BF^{-1}[C_1 \mid C_2] \begin{bmatrix} F & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \\ &= B[F^{-1}C_1F \mid F^{-1}C_2]. \end{aligned} \quad (\text{C4})$$

Meanwhile, with A also in block matrix form, i.e.,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = [A_1 \mid A_2], \quad (\text{C5})$$

where A_1 and A_2 are of size $L_s \times r$ and $L_s \times (N_s - r)$, respectively, A_{11} and A_{21} are of size of $r \times r$ and $(L_s - r) \times r$, respectively, etc, we have

$$\begin{aligned} A(\tilde{F} - I) &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} F - I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}(F - I) & \mathbf{0} \\ A_{21}(F - I) & \mathbf{0} \end{bmatrix} \\ &= A_1[(F - I) \quad \mathbf{0}]. \end{aligned} \quad (\text{C6})$$

Combining Eq. (C4) and Eq. (C6) together, we eventually obtain

$$\begin{aligned} \Psi_{\text{HF}}(c) &= \det \begin{bmatrix} A[c] & B[c] \\ C_\theta(c) & F_\theta(c) \end{bmatrix} \\ &= \det((A - B'C'_\theta(c))[c]), \end{aligned} \quad (\text{C7})$$

where

$$B' = [-A_1 \quad B], \quad C'_\theta(c) = \begin{bmatrix} F - I & \mathbf{0} \\ F^{-1}C_1F & F^{-1}C_2 \end{bmatrix} \quad (\text{C8})$$

are of size $L_s \times 2r$ and $2r \times N_s$, respectively.

Likewise, for r -JBF, we have

$$\begin{aligned} \Psi_{\text{JBF}}(c) &= \det(F_\theta(c)) \det((A - BC_\theta(c))[c]) \\ &= \det((A - B'C'_\theta(c))[c]), \end{aligned} \quad (\text{C9})$$

where

$$B' = [-A_1 \quad B], \quad C'_\theta(c) = \begin{bmatrix} F - I & \mathbf{0} \\ C_1F & C_2 \end{bmatrix} \quad (\text{C10})$$

are of size $L_s \times 2r$ and $2r \times N_s$, respectively.

APPENDIX D: RELATION BETWEEN WAVE-FUNCTION FIDELITY AND SPO DISTANCE

In Fig. 5, we observe that at $r = 1, 2$, though the SPO distance Eq. (13) is small [Fig. 5(a)], the wave-function fidelity Eq. (12) is relatively worse [Fig. 5(b)], compared with larger r cases. In this section, we will look into this observation by showing the relation between the wave-function fidelity and the SPO distance.

Within VMC, the wave-function fidelity can be calculated as

$$\mathcal{F}_{\text{wf}} = \frac{\left(\frac{1}{S} \sum_{\{c\}} \frac{\Phi(c)}{\Psi(c)} \right)^2}{\frac{1}{S} \sum_{\{c\}} \left(\frac{\Phi(c)}{\Psi(c)} \right)^2}, \quad (\text{D1})$$

where S is the size of sample set $\{c\}$ sampled from $|\Psi(c)|^2$. As $\Phi(c)/\Psi(c)$ is close to 1 here, we rewrite it as

$$\frac{\Phi(c)}{\Psi(c)} \equiv 1 + x(c). \quad (\text{D2})$$

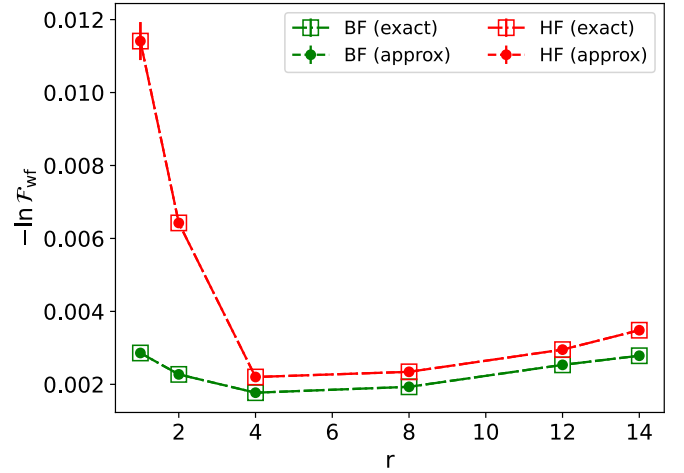


FIG. 9. Comparison between the approximate results Eq. (D6) for the wave-function fidelity and the exact results from VMC. The target wave functions are r -JBF from ground state approximation, and $2r$ -BF and r -HF are optimized towards the target wave functions by minimizing the SPO distances.

For most of c , we should have $|x(c)| \ll 1$. We relate $x(c)$ to the SPO distance component with Taylor expansion around $\Psi(c)$ up to the first order,

$$x(c) = \frac{\det \bar{A}(c)}{\det \bar{A}_0(c)} - 1 \approx \sum_{jk} \bar{A}_0(c)^{-1}_{kj} (\bar{A}(c) - \bar{A}_0(c))_{jk}, \quad (\text{D3})$$

where $\bar{A}(c)$ and $\bar{A}_0(c)$ are the square matrices in the determinant wave-function form of $|\Phi\rangle$ and $|\Psi\rangle$, respectively. Note that the truncation in Taylor expansion is one of the sources of error for our final approximation formula, which can nonetheless be systemically improved by including higher-order terms.

In the following, we will perform Taylor expansion up to the second order on wave-function fidelity as well. The first-order term will vanish as $\mathcal{F}_{\text{wf}} = 1$ is a stationary point. After defining

$$\bar{x} = \frac{1}{S} \sum_{\{c\}} x(c), \quad \bar{x}^2 = \frac{1}{S} \sum_{\{c\}} x(c)^2 \quad (\text{D4})$$

and $|\bar{x}| \ll 1$, $|\bar{x}^2| \ll 1$, we can either use chain rule to obtain the Hessian as

$$\frac{\delta}{\delta \bar{A}(c')_{j'k'}} \frac{\delta \ln \mathcal{F}_{\text{wf}}}{\delta \bar{A}(c)_{jk}} = \frac{1}{S^2} \bar{A}_0(c)^{-1}_{kj} \bar{A}_0(c')^{-1}_{k'j'} (1 - S \delta_{c,c'}) \quad (\text{D5})$$

or directly expand on $\ln \mathcal{F}_{\text{wf}}$ in terms of \bar{x} and \bar{x}^2 , we obtain the approximation formula for $\ln \mathcal{F}_{\text{wf}}$, the deviation from perfect fidelity $\mathcal{F}_{\text{wf}} = 1$, as the variance of $x(c)$,

$$\ln \mathcal{F}_{\text{wf}} \approx \bar{x}^2 - \bar{x}^2. \quad (\text{D6})$$

In Fig. 9, we examine the accuracy of Eq. (D6) by comparing with the exact results from VMC.

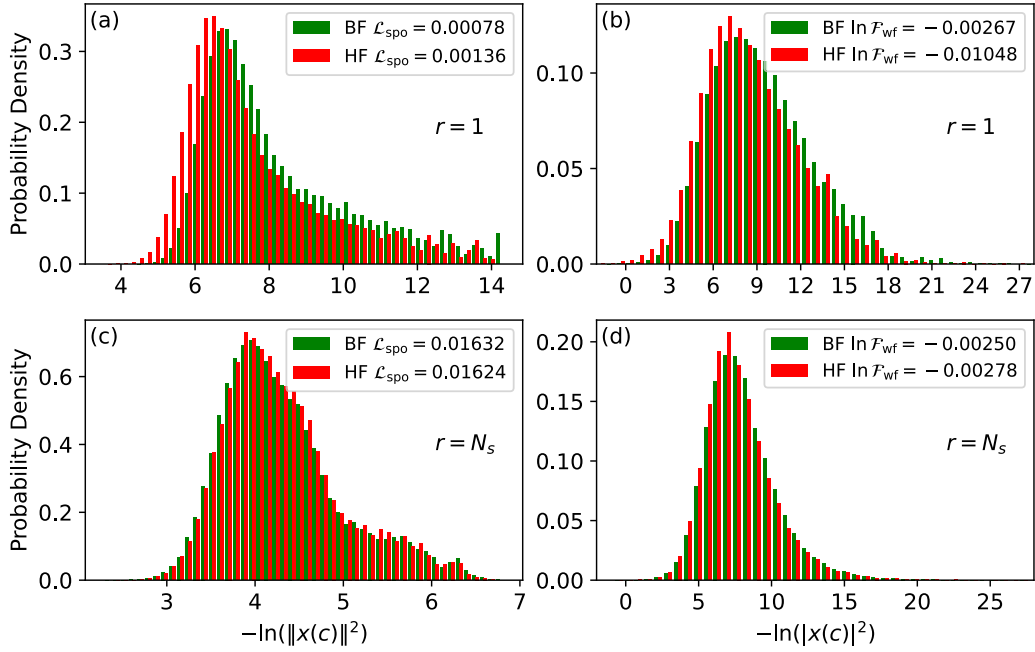


FIG. 10. Probability distributions of $\|x(c)\|^2$ and $|x(c)|^2$ at $r = 1$ and $14(N_s)$. The corresponding SPO distance and wave-function fidelity are given in the legends.

We also found that $\bar{x}^2 \ll \overline{x^2}$. Then by just comparing \bar{x}^2 with the SPO distance Eq. (13), which we rewrite as

$$\mathcal{L}_{\text{spo}} = \frac{1}{S} \sum_{\{c\}} \|x(c)\|^2 = \frac{1}{S} \sum_{\{c\}} \sum_{jk} |\bar{A}(c) - \bar{A}_0(c)|_{jk}^2, \quad (\text{D7})$$

we can see that wave-function fidelity is effectively a weighted average of SPO distance on each element, with the weights given by $\bar{A}_0(c)^{-1}$ from the target state. Further evidence is shown from the comparison between Figs. 10(a), 10(c) and 10(b), 10(d) where the difference of wave-function fidelity is affected not only by the SPO distance, but also the target wave function.

APPENDIX E: RECONSTRUCTION OF r -JBF WITH $2r$ -BF

Despite the existence of matrix multiplication within r -JBF, it is possible to reconstruct r -JBF rigorously from $2r$ -BF with feed-forward neural networks still. In this section we show it explicitly.

First, with the feed-forward neural network that generates backflow matrices $C_\theta(c)$ and $F_\theta(c)$ for r -JBF, we add an activation function $a(x) = \ln(x)$ to this layer (note that originally for r -JBF, this layer does not have activation function on it. Also note that for the case where $x = 0$, we can nonetheless replace it with $x + \epsilon$, $\epsilon \ll 1$, which will not affect the subsequent steps). Now, we have the output from this modified layer as $\ln C_{jk}$ ($j = 1, 2, \dots, r$; $k = 1, 2, \dots, N_s$), and $\ln F_{lm}$ ($l, m = 1, 2, \dots, r$). In the following, we will work on the first r columns of $\ln C$, i.e., $\ln C_1$, in order to reproduce $C_1 F$ with feed-forward neural network. As for the remaining $\ln C_2$, we construct the weights and bias as an identity map, and the nonlinear activation function introduced later will bring $\ln C_2$ back to the original C_2 from r -JBF.

Next, we add one more layer to this neural network, whose activation function is $a(x) = \exp(x)$, bias are 0's, and

weights are

$$W_{(xyz)(jk)}^{[C]} = \delta_{xj} \delta_{zk}, \quad W_{(xyz)(lm)}^{[F]} = \delta_{zl} \delta_{ym}, \quad (\text{E1})$$

where $x, y, z = 1, 2, \dots, r$, such that the output from this layer is

$$\begin{aligned} D_{xyz} &= \exp \left(\sum_{jk} W_{(xyz)(jk)}^{[C]} \ln C_{1,jk} + \sum_{lm} W_{(xyz)(lm)}^{[F]} \ln F_{lm} \right) \\ &= C_{1,xz} F_{zy}. \end{aligned} \quad (\text{E2})$$

At the end, we add one more linear layer with nonzero weights as

$$W_{(jm)(xyz)} = \delta_{jx} \delta_{my} \quad (\text{E3})$$

such that the output is

$$\tilde{C}_1 = \sum_{xyz} W_{(jm)(xyz)} D_{xyz} = \sum_z C_{1,jz} F_{zm} \equiv (C_1 F)_{jm}. \quad (\text{E4})$$

Overall, by introducing two more layers of widths $r^2(N_s + r)$ and $r(N_s + r)$ to the original neural network for r -JBF, we are able to reproduce the backflow matrices with $2r$ -BF. Therefore we have analytically proved that r -JBF is a subset of r -BF at finite w if we allow r -BF to have additional layers.

APPENDIX F: COMPUTATIONAL COST

In this section, we give the computational cost for obtaining the amplitude from various types of NNBF wave functions.

Let us start with NN part. The input size is L_s , and the output size is $rN_s + r^2$, and assume the width of NN is w , then

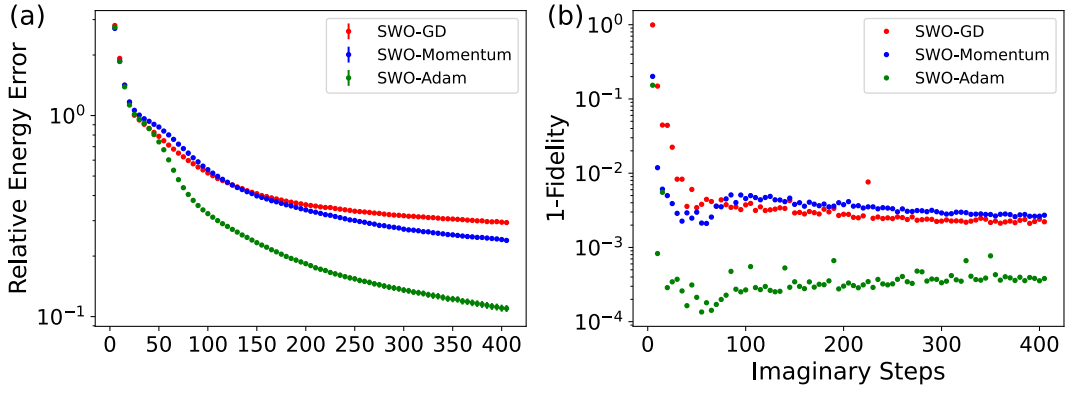


FIG. 11. (a) Energy and (b) fidelity [associated with the loss function Eq. (8)] from r -HF at $r = 64$ during SWO optimization ($\tau = 0.018$) with optimizer SGD (red), momentum (blue), and Adam (green) for Hubbard model ($L = 4 \times 8$, PBC, $U/t = 8$, $N_\uparrow = N_\downarrow = 14$).

the computational cost due to the matrix-vector multiplication is given by $O(w^2 + wL_s + wr^2 + wrN_s)$.

Next, the cost of the evaluation of determinant is

$$\begin{aligned} r\text{-HF} &: O((N_s + r)^3) = O(N_s^3 + r^3), \\ r\text{-BF} &: O(rN_s^2 + N_s^3), \\ r\text{-JBF} &: O(r^3 + rN_s^2 + N_s^3). \end{aligned} \quad (\text{F1})$$

In Appendix E, we show that to reproduce r -JBF from $2r$ -BF, we can add two more layers of width $r^3 + r^2N_s$ and $rN_s + r^2$, this will give an extra cost as $O(r^3(N_s + r)^2) = O(r^3N_s^2 + r^5)$.

APPENDIX G: FAILURE OF SPO MATCHING WITH JSD TOWARDS NNBF

In this section, we illustrate the impossibility of matching the SPO of NNBF with JSD for rank $r \leq N_s$, thus, combining with the numerical evidences Figs. 2(b) and 8, we conclude that NNBF represents a broader space of SPO.

For simplicity, we consider the case of N_s -JSD and N_s -BF. Since the NN generating the back flow matrices has linear layer as the last layer, there are constant matrices b in the backflow matrices, then we rewrite the SPO as

$$\begin{aligned} S_{\text{JSD}}(c) &= A_{\text{JSD}} F_{\text{JSD}, \theta}(c) \\ &= A_{\text{JSD}} b_{\text{JSD}} + A_{\text{JSD}} \tilde{F}_{\text{JSD}, \theta}(c), \\ S_{\text{BF}}(c) &= A_{\text{BF}} - B_{\text{BF}} C_{\text{BF}, \theta}(c) \\ &= A_{\text{BF}} - B_{\text{BF}} b_{\text{BF}} - B_{\text{BF}} \tilde{C}_{\text{BF}, \theta}(c). \end{aligned} \quad (\text{G1})$$

In order to match the SPO, $S_{\text{JSD}}(c) = S_{\text{BF}}(c)$, we have to solve the following equations for static matrices:

$$\begin{aligned} A_{\text{JSD}} &= -B_{\text{BF}}, \\ A_{\text{JSD}} b_{\text{JSD}} &= A_{\text{BF}} - B_{\text{BF}} b_{\text{BF}}. \end{aligned} \quad (\text{G2})$$

However, the second equation above is an over-determinant equation, so there is no exact solution for b_{JSD} , unless $A_{\text{BF}} - B_{\text{BF}} b_{\text{BF}}$ falls into the linear space spanned by the column vectors of A_{JSD} . However, this special case is not generically true for the BF wave function we are considering.

In conclusion, we have showed that it is impossible for JSD to reproduce the same SPO from BF.

APPENDIX H: SUPERVISED WAVE-FUNCTION OPTIMIZATION

In this section, we provide detailed information on the SWO method, including the derivation for the gradient formula (9), we use for NNBF (and its variants) and comparison between various optimizers.

To begin with, we note that SWO is a first-order optimization method for variational wave function to undergo imaginary time evolution in a stochastic way. As already mentioned in Sec. IV, at each imaginary-time step t , we set the target state as the state evolved by imaginary time τ , i.e., $|\Phi_t\rangle = e^{-\tau \hat{H}} |\Psi_{t-1}\rangle \approx (1 - \tau \hat{H}) |\Psi_{t-1}\rangle$, where τ needs to be small enough for the latter approximation to be valid. In practice, we notice that the energy can be minimized successfully with $\tau < 1/E_{t=0}$, where $E_{t=0}$ is the variational energy of randomly initialized state.

To derive the gradient formula Eq. (9), we first take derivative of the loss function Eq. (8) with respect to the parameter θ ,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta} &= -\frac{\partial_\theta \langle \Psi_t | \Phi_t \rangle}{\langle \Psi_t | \Phi_t \rangle} - \frac{\partial_\theta \langle \Phi_t | \Psi_t \rangle}{\langle \Phi_t | \Psi_t \rangle} + \frac{\partial_\theta \langle \Psi_t | \Psi_t \rangle}{\langle \Psi_t | \Psi_t \rangle} \\ &= -\frac{\sum_c (\partial_\theta \langle \Psi_t | c \rangle) \langle c | \Phi_t \rangle}{\sum_c \langle \Psi_t | c \rangle \langle c | \Phi_t \rangle} - \frac{\sum_c \langle \Phi_t | c \rangle (\partial_\theta \langle c | \Psi_t \rangle)}{\sum_c \langle \Phi_t | c \rangle \langle c | \Psi_t \rangle} \\ &\quad + \frac{\sum_c (\partial_\theta \langle \Psi_t | c \rangle) \langle c | \Psi_t \rangle + \sum_c \langle \Psi_t | c \rangle (\partial_\theta \langle c | \Psi_t \rangle)}{\sum_c \langle \Psi_t | c \rangle \langle c | \Psi_t \rangle}. \end{aligned} \quad (\text{H1})$$

This expression is exact since we are summing over all the configurations c . Next, we use the sample set $\{c|p(c)\}$ from probability distribution $p(c) = |\Psi_{t-1}(c)|^2$ to replace the exact summation. Then, for example, the first term of Eq. (H1) ends up as

$$\begin{aligned} \frac{\sum_c (\partial_\theta \langle \Psi_t | c \rangle) \langle c | \Phi_t \rangle}{\sum_c \langle \Psi_t | c \rangle \langle c | \Phi_t \rangle} &= \frac{\sum_c \frac{(\partial_\theta \langle \Psi_t | c \rangle) \langle c | \Phi_t \rangle}{p(c)} p(c)}{\sum_c \frac{\langle \Psi_t | c \rangle \langle c | \Phi_t \rangle}{p(c)} p(c)} \\ &\approx \frac{\sum_{\{c\}} \frac{(\partial_\theta \langle \Psi_t | c \rangle) \langle c | \Phi_t \rangle}{p(c)}}{\sum_{\{c\}} \frac{\langle \Psi_t | c \rangle \langle c | \Phi_t \rangle}{p(c)}} \\ &= \frac{\sum_{\{c\}} \frac{(\partial_\theta \langle \Psi_t | c \rangle) \langle c | \Phi_t \rangle}{\langle \Psi_{t-1} | c \rangle \langle c | \Psi_{t-1} \rangle}}{\sum_{\{c\}} \frac{\langle \Psi_t | c \rangle \langle c | \Phi_t \rangle}{\langle \Psi_{t-1} | c \rangle \langle c | \Psi_{t-1} \rangle}}. \end{aligned} \quad (\text{H2})$$

As we are optimizing over determinant wave-function form, i.e., $\Psi_\theta(c) = \det \tilde{A}_\theta(c)$, we can further simplify the derivative term in Eq. (H1) as

$$\begin{aligned} \frac{\partial \Psi(c)}{\partial \theta} &= \frac{\partial \det \tilde{A}_\theta(c)}{\partial \theta} = \det \tilde{A}_\theta(c) \text{Tr} \left(\tilde{A}_\theta(c)^{-1} \frac{\partial \tilde{A}_\theta(c)}{\partial \theta} \right) \\ &= \Psi(c) \text{Tr} \left(\tilde{A}_\theta(c)^{-1} \frac{\partial \tilde{A}_\theta(c)}{\partial \theta} \right). \end{aligned} \quad (\text{H3})$$

After defining the amplitude ratio $\phi(c)$ and $\psi(c)$ as Eq. (11) to simplify the notations, we factor out the derivative part, and put together all the other factors as a coefficient $\alpha(c)$ given by Eq.(10), then we end up with the gradient estimation formula Eq. (9). That only the real part of gradient is taken is because the parameter set θ are real numbers and each term in Eq. (H1) is accompanied with its complex conjugate.

Finally, we note that, as SWO is a first-order optimization method, it can be further improved with ML optimizers, such as Momentum-based gradient descent, Adam, etc. In Fig. 11, we compare the performance from gradient descent (GD), momentum, and Adam within SWO on the ground state approximation with r -HF. It turns out that SWO-Adam is more efficient than other two optimizers, as it optimizes the fidelity to a higher value at each step, so as to minimize the

energy faster than the other two optimizers. Henceforth, we use SWO-Adam for optimization in our numerical tests.

APPENDIX I: NOTATIONS

In Table I, we give a summary for the notations used in this paper.

TABLE I. Table of Notations.

c	L_s -size binary vector
θ	Parameters (weights and bias) of neural networks
A	$L_s \times N_s$ constant matrix
$A[c]$	$N_s \times N_s$ matrix from row selection on A based on c
B	$L_s \times r$ constant matrix
$B[c]$	$N_s \times r$ matrix from row selection on B based on c
$C_\theta(c)$	$r \times N_s$ backflow matrix from neural networks,
$F_\theta(c)$	$r \times r$ backflow matrix from neural networks,
r -HF	$\Psi(c) = \det(F_\theta(c)) \det((A - BF_\theta(c))^{-1} C_\theta(c))[c]$
r -JBF	$\Psi(c) = \det(F_\theta(c)) \det((A - BC_\theta(c))[c])$
r -BF	$\Psi(c) = \det((A - BC_\theta(c))[c])$
r -JNNBF	$\Psi(c) = \det(F_\theta(c)) \Psi_{L_s\text{-BF}}(c)$
r -JSD	$\Psi(c) = \det(F_\theta(c)) \det A[c]$
r -expJBF	$\Psi(c) = \exp(f_\theta(c)) \Psi_{r\text{-BF}}(c)$
r -expJsBF	$\Psi(c) = \exp(f_{1,\theta}(c)) \cos(f_{2,\theta}(c)) \Psi_{r\text{-BF}}(c)$

- [1] J. C. Slater, The theory of complex spectra, *Phys. Rev.* **34**, 1293 (1929).
- [2] R. Jastrow, Many-body problem with strong forces, *Phys. Rev.* **98**, 1479 (1955).
- [3] L. Mitáš and R. M. Martin, Quantum Monte Carlo of nitrogen: Atom, dimer, atomic, and molecular solids, *Phys. Rev. Lett.* **72**, 2438 (1994).
- [4] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Quantum Monte Carlo simulations of solids, *Rev. Mod. Phys.* **73**, 33 (2001).
- [5] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
- [6] 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).
- [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] A. Lovato, C. Adams, G. Carleo, and N. Rocco, Hidden-nucleons neural-network quantum states for the nuclear many-body problem, *Phys. Rev. Res.* **4**, 043178 (2022).
- [9] K. Choo, A. Mezzacapo, and G. Carleo, Fermionic neural-network states for *ab-initio* electronic structure, *Nat. Commun.* **11**, 2368 (2020).
- [10] N. Yoshioka, W. Mizukami, and F. Nori, Solving quasiparticle band spectra of real solids using neural-network quantum states, *Commun. Phys.* **4**, 106 (2021).
- [11] D. Luo and B. K. Clark, Backflow transformations via neural networks for quantum many-body wave functions, *Phys. Rev. Lett.* **122**, 226401 (2019).
- [12] 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).
- [13] J. Hermann, Z. Schätzle, and F. Noé, Deep-neural-network solution of the electronic Schrödinger equation, *Nat. Chem.* **12**, 891 (2020).
- [14] J. S. Spencer, D. Pfau, A. Botev, and W. M. C. Foulkes, Better, faster fermionic neural networks, [arXiv:2011.07125](https://arxiv.org/abs/2011.07125).
- [15] K. Inui, Y. Kato, and Y. Motome, Determinant-free fermionic wave function using feed-forward neural networks, *Phys. Rev. Res.* **3**, 043126 (2021).
- [16] 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).
- [17] J. Stokes, J. R. Moreno, E. A. Pnevmatikakis, and G. Carleo, Phases of two-dimensional spinless lattice fermions with first-quantized deep neural-network quantum states, *Phys. Rev. B* **102**, 205122 (2020).
- [18] F. Ferrari, F. Becca, and J. Carrasquilla, Neural Gutzwiller-projected variational wave functions, *Phys. Rev. B* **100**, 125131 (2019).
- [19] D. A. Huse and V. Elser, Simple variational wave functions for two-dimensional Heisenberg spin-1/2 antiferromagnets, *Phys. Rev. Lett.* **60**, 2531 (1988).

- [20] H. J. Changlani, J. M. Kinder, C. J. Umrigar, and G. K.-L. Chan, Approximating strongly correlated wave functions with correlator product states, *Phys. Rev. B* **80**, 245116 (2009).
- [21] F. Mezzacapo, N. Schuch, M. Boninsegni, and J. I. Cirac, Ground-state properties of quantum many-body systems: Entangled-plaquette states and variational Monte Carlo, *New J. Phys.* **11**, 083026 (2009).
- [22] R. J. Needs, M. D. Towler, N. D. Drummond, P. L. Ríos, and J. R. Trail, Variational and diffusion quantum Monte Carlo calculations with the CASINO code, *J. Chem. Phys.* **152**, 154106 (2020).
- [23] N. Abrahamsen and L. Lin, Anti-symmetric Barron functions and their approximation with sums of determinants, [arXiv:2303.12856](https://arxiv.org/abs/2303.12856).
- [24] M. Ruggeri, S. Moroni, and M. Holzmann, Nonlinear network description for many-body quantum systems in continuous space, *Phys. Rev. Lett.* **120**, 205302 (2018).
- [25] Z. Chen, D. Luo, K. Hu, and B. K. Clark, Simulating 2+1D lattice quantum electrodynamics at finite density with neural flow wavefunctions, [arXiv:2212.06835](https://arxiv.org/abs/2212.06835).
- [26] E. Dagotto, A. Moreo, F. Ortolani, D. Poilblanc, and J. Riera, Static and dynamical properties of doped Hubbard clusters, *Phys. Rev. B* **45**, 10741 (1992).
- [27] D. Kochkov and B. K. Clark, Variational optimization in the AI era: Computational graph states and supervised wave function optimization, [arXiv:1811.12423](https://arxiv.org/abs/1811.12423).
- [28] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [29] A. Pinkus, Approximation theory of the MLP model in neural networks, *Acta Numerica* **8**, 143 (1999).
- [30] S. R. White, Density matrix formulation for quantum renormalization groups, *Phys. Rev. Lett.* **69**, 2863 (1992).
- [31] F. Verstraete, M. M. Wolf, D. Perez-Garcia, and J. I. Cirac, Criticality, the area law, and the computational power of projected entangled pair states, *Phys. Rev. Lett.* **96**, 220601 (2006).
- [32] G. Vidal, Entanglement renormalization, *Phys. Rev. Lett.* **99**, 220405 (2007).
- [33] R. Orús, Tensor networks for complex quantum systems, *Nat. Rev. Phys.* **1**, 538 (2019).
- [34] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, *Ann. Phys.* **349**, 117 (2014).
- [35] G. Vidal, Class of quantum many-body states that can be efficiently simulated, *Phys. Rev. Lett.* **101**, 110501 (2008).
- [36] D.-L. Deng, X. Li, and S. Das Sarma, Quantum entanglement in neural network states, *Phys. Rev. X* **7**, 021021 (2017).
- [37] G. Passetti, D. Hofmann, P. Neitemeier, L. Grunwald, M. A. Sentef, and D. M. Kennes, Can neural quantum states learn volume-law ground states? *Phys. Rev. Lett.* **131**, 036502 (2023).
- [38] Z. Denis, A. Sinibaldi, and G. Carleo, Comment on “can neural quantum states learn volume-law ground states?”, [arXiv:2309.11534](https://arxiv.org/abs/2309.11534).
- [39] F. B. Trigueros, T. Mendes-Santos, and M. Heyl, Simplicity of mean-field theories in neural quantum states, *Phys. Rev. Res.* **6**, 023261 (2024).
- [40] O. Sharir, A. Shashua, and G. Carleo, Neural tensor contractions and the expressive power of deep neural quantum states, *Phys. Rev. B* **106**, 205136 (2022).
- [41] S. R. Clark, Unifying neural-network quantum states and correlator product states via tensor networks, *J. Phys. A: Math. Theor.* **51**, (2018).